



AI Guided Ontology Curation Workflows

<https://bit.ly/ai-ontology-workflows>

Chris Mungall
Lawrence Berkeley National Laboratory

OBO Academy
March 2024

Goals for today

- Demo / Tutorial
 - Demonstrate what can be done using standard **generic** ChatGPT (and similar) interfaces
 - Next session we will talk about **specialized** tools (OntoGPT, CurateGPT, OAK-LLM)
- You can follow along if you want
 - These slides: <https://bit.ly/ai-ontology-workflows>
- Questions encouraged!
- Goal is to make you feel empowered to experiment and **explore for yourself**
- If we have time
 - Report on evaluation of DRAGON-AI

Caveat 1: AI is weird

- Everything I aim to show today has worked in the past
- But web-based AIs are non-deterministic
- It's quite possible that things will go wrong – this is an experiment!



Caveat 2: Tools are not open to all

- ChatGPT
 - <https://chat.openai.com/>
 - Gpt-4 requires subscription
- Gemini
 - <https://gemini.google.com/app>
 - AI premium account
- Claude
 - <https://claude.ai/chats>
 - Claude-3 Opus requires a subscription
- Groq
 - <https://groq.com/>
 - Free (for now!)
 - Note: this is nothing to do with Twitter's Grok
 - Fast access to mixtral and llama2

Experiments today

- Direct chat interface
- Review-based workflows
- Enhancement workflows
- ROBOT helper
- Consensus for finding references

First attempt - obo format

Create a term in the plant ontology for the concept "Fine root C/leaf C ratio". Show in obo format

<https://chat.openai.com>

First attempt - seems OK?



You

create a term in the plant trait ontology for the concept "Fine root C/leaf C ratio". Show in obo format



ChatGPT

To create a term in the Plant Trait Ontology (PTO) for the concept "Fine root C/leaf C ratio" in OBO (Open Biomedical Ontologies) format, we need to follow the standard conventions of OBO format. Here's an example of how this new term could be represented:

```
vbnet Copy code
[Term]
id: PT0:1234567
name: fine root C/leaf C ratio
namespace: plant_trait_ontology
def: "The ratio of carbon content in fine roots to the carbon content in l
synonym: "root to leaf carbon ratio" EXACT []
is_a: PT0:0000002 ! plant anatomical entity property
relationship: part_of PT0:0000203 ! root anatomy and morphology trait
relationship: part_of PT0:0000212 ! leaf anatomy and morphology trait
```



Uh oh! Hallucinations!



You

create a term in the plant trait ontology for the concept "Fine root C/leaf C ratio". Show in obo format

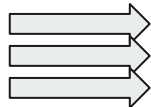


ChatGPT

To create a term in the Plant Trait Ontology (PTO) for the concept "Fine root C/leaf C ratio" in OBO (Open Biomedical Ontologies) format, we need to follow the standard conventions of OBO format. Here's an example of how this new term could be represented:

```
vbnet Copy code
[Term]
id: PT0:1234567
name: fine root C/leaf C ratio
namespace: plant_trait_ontology
def: "The ratio of carbon content in fine roots to the carbon content in l
synonym: "root to leaf carbon ratio" EXACT []
is_a: PT0:0000002 ! plant anatomical entity property
relationship: part_of PT0:0000203 ! root anatomy and morphology trait
relationship: part_of PT0:0000212 ! leaf anatomy and morphology trait
```

No such term



Too general a parent
(ChatGPT doesn't know the ontology well enough to place a term)



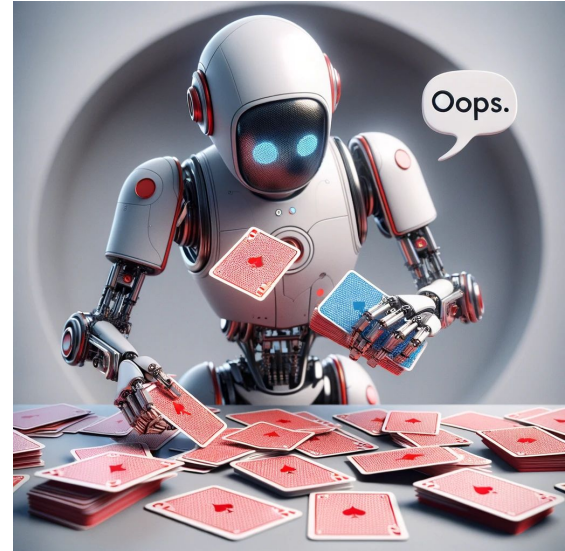
ROBOT CSVs to the rescue

- ROBOT CSVs are nice for humans as we don't have to memorize IDs
- Nice for AIs for the same reason!

| | A | B | C | D | E | F | G | H | I | J | N | O | P | Q |
|----|-------------|---------------------------------|-------------------|--------------|-----------------|---------------------------------------------------------|-----------------------|-----------------|----------|--------------|--------------------------------------------------------------------------|----------------------|------------------------------|----|
| 1 | ontology ID | label | editor prefer | has curation | alternative t | definition | definition so exam | term ed | editor n | logical type | specimen collection proc | gross anatomical p | collection process input | UF |
| 2 | ID | A rdfs:label | A editor pref AI | has curati | A alternative A | A definition | A definition : A exar | A term € | A editor | CLASS_TYPE | C specimen and (is_speci | C specimen and (is_C | specimen and (is_specified_t | |
| 3 | OBI:0002500 | amniotic fluid specimen | ready for release | | | A specimen that is derived from amniotic fluid. | Chris Stoeckert, Pen | Chris Stoeckert | | equivalent | collecting specimen from amniotic fluid | | | UF |
| 4 | OBI:0002501 | bile specimen | ready for release | | | A specimen that is derived from bile. | Chris Stoeckert, Pen | Chris Stoeckert | | equivalent | collecting specimen from bile | | | UF |
| 5 | OBI:0002502 | cerebrospinal fluid specimen | ready for release | | | A specimen that is derived from cerebrospinal fluid. | Chris Stoeckert, Pen | Chris Stoeckert | | equivalent | collecting specimen from cerebrospinal fluid | | | UF |
| 6 | OBI:0002503 | feces specimen | ready for release | | | A specimen that is derived from feces. | Chris Stoeckert, Pen | Chris Stoeckert | | equivalent | collecting specimen from feces | | | UF |
| 7 | OBI:0002504 | digestive system fluid or s | ready for release | | | A specimen that is derived from digestive system fluid. | Chris Stoeckert, Pen | Chris Stoeckert | | equivalent | collecting specimen from digestive system secreted substance | | | UF |
| 8 | OBI:0002505 | milk specimen | ready for release | | | A specimen that is derived from milk. | Chris Stoeckert, Pen | Chris Stoeckert | | equivalent | collecting specimen from milk | | | UF |
| 9 | OBI:0002506 | pericardial fluid specimen | ready for release | | | A specimen that is derived from pericardial fluid. | Chris Stoeckert, Pen | Chris Stoeckert | | equivalent | collecting specimen from pericardial fluid | | | UF |
| 10 | OBI:0002507 | saliva specimen | ready for release | | | A specimen that is derived from saliva. | Chris Stoeckert, Pen | Chris Stoeckert | | equivalent | collecting specimen from saliva | | | UF |
| 11 | OBI:0002508 | sputum specimen | ready for release | | | A specimen that is derived from sputum. | Chris Stoeckert, Pen | Chris Stoeckert | | equivalent | collecting specimen from sputum | | | UF |
| 12 | OBI:0002509 | sweat specimen | ready for release | | | A specimen that is derived from sweat. | Chris Stoeckert, Pen | Chris Stoeckert | | equivalent | collecting specimen from sweat | | | UF |
| 13 | OBI:0002510 | synovial fluid specimen | ready for release | | | A specimen that is derived from synovial fluid. | Chris Stoeckert, Pen | Chris Stoeckert | | equivalent | collecting specimen from synovial fluid | | | UF |
| 14 | OBI:0002511 | vitreous humor specimen | ready for release | | | A specimen that is derived from vitreous humor. | Chris Stoeckert, Pen | Chris Stoeckert | | equivalent | collecting specimen from (tissue and ('part of' some 'vitreous humor')) | | | UF |
| 15 | OBI:0002512 | bone marrow specimen | ready for release | | | A specimen that is derived from bone marrow. | Chris Stoeckert, Pen | Chris Stoeckert | | equivalent | bone marrow biopsy bone marrow | | | UF |
| 16 | OBI:0002513 | placenta specimen | ready for release | | | A specimen that is derived from placenta. | Chris Stoeckert, Pen | Chris Stoeckert | | equivalent | collecting specimen from (tissue and ('part of' some 'placenta')) | | | UF |
| 17 | OBI:0002514 | peritoneal fluid specimen | ready for release | | | A specimen that is derived from peritoneal fluid. | Chris Stoeckert, Pen | Chris Stoeckert | | equivalent | collecting specimen from peritoneal fluid | | | UF |
| 18 | OBI:0002515 | pleural fluid specimen | ready for release | | | A specimen that is derived from pleural fluid. | Chris Stoeckert, Pen | Chris Stoeckert | | equivalent | collecting specimen from pleural fluid | | | UF |
| 19 | OBI:0002516 | brain specimen | ready for release | | | A specimen that is derived from brain. | Chris Stoeckert, NCI | Chris Stoeckert | | equivalent | collecting specimen from (tissue and ('part of' some 'brain')) | | | UF |
| 20 | OBI:0002517 | hair specimen | ready for release | | | A specimen that is derived from hair. | Chris Stoeckert, NCI | Chris Stoeckert | | equivalent | collecting specimen from strand of hair | | | UF |
| 21 | OBI:0002518 | prostate gland specimen | ready for release | | | A specimen that is derived from prostate gland. | Chris Stoeckert, NCI | Chris Stoeckert | | equivalent | collecting specimen from (tissue and ('part of' some 'prostate gland')) | | | UF |
| 22 | OBI:0002519 | skeletal muscle tissue specimen | ready for release | | | A specimen that is derived from skeletal muscle. | Chris Stoeckert, NCI | Chris Stoeckert | | equivalent | collecting specimen from skeletal muscle tissue | | | UF |
| 23 | OBI:0002520 | heart specimen | ready for release | | | A specimen that is derived from heart. | Chris Stoeckert, NCI | Chris Stoeckert | | equivalent | collecting specimen from (tissue and ('part of' some 'heart')) | | | UF |
| 24 | OBI:0002521 | renal medulla specimen | ready for release | | | A specimen that is derived from renal medulla. | Chris Stoeckert, NCI | Chris Stoeckert | | equivalent | collecting specimen from (tissue and ('part of' some 'renal medulla')) | | | UF |
| 25 | OBI:0002522 | adrenal gland specimen | ready for release | | | A specimen that is derived from adrenal gland. | Chris Stoeckert, NCI | Chris Stoeckert | | equivalent | collecting specimen from (tissue and ('part of' some 'adrenal gland')) | | | UF |
| 26 | OBI:0002523 | breast specimen | ready for release | | | A specimen that is derived from breast. | Chris Stoeckert, NCI | Chris Stoeckert | | equivalent | collecting specimen from (tissue and ('part of' some 'breast')) | | | UF |
| 27 | OBI:0002524 | urinary bladder specimen | ready for release | | | A specimen that is derived from urinary bladder. | Chris Stoeckert, NCI | Chris Stoeckert | | equivalent | collecting specimen from (tissue and ('part of' some 'urinary bladder')) | | | UF |
| 28 | OBI:0002525 | tibial artery specimen | ready for release | | | A specimen that is derived from tibial artery. | Chris Stoeckert, NCI | Chris Stoeckert | | equivalent | collecting specimen from (tissue and ('part of' some 'tibial artery')) | | | UF |
| 29 | OBI:0002526 | skin of body specimen | ready for release | | | A specimen that is derived from skin. | Chris Stoeckert, NCI | Chris Stoeckert | | equivalent | collecting specimen from (tissue and ('part of' some 'skin of body')) | | | UF |
| 30 | OBI:0002527 | pancreas specimen | ready for release | | | A specimen that is derived from pancreas. | Chris Stoeckert, NCI | Chris Stoeckert | | equivalent | collecting specimen from (tissue and ('part of' some 'pancreas')) | | | UF |
| 31 | OBI:0002528 | stomach specimen | ready for release | | | A specimen that is derived from stomach. | Chris Stoeckert, NCI | Chris Stoeckert | | equivalent | collecting specimen from (tissue and ('part of' some 'stomach')) | | | UF |

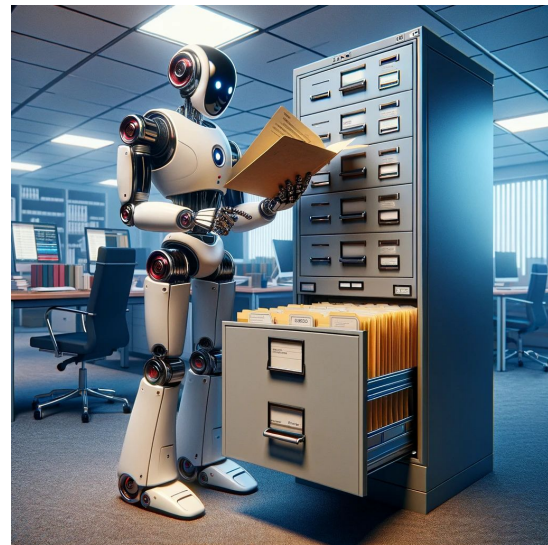
Review-based workflow

- It's easy to make mistakes with ontologies!
 - Hard to avoid these slipping through



Review-based workflow

- It's easy to make mistakes with ontologies!
 - Hard to avoid these slipping through
- Let's have the AI review a collection of ontology terms
- The collection will be represented as a robot template CSV
- We will introduce a deliberate error
 - <https://gist.github.com/cmungall/68ceedb93abdd24946cb2e3737bbe603>
 - Can you find the error faster than the AI?
- Let's try ChatGPT



Review-based workflow

Ontology Terms Structured Format

March 4, 2024



Anonymous

Check that the definitions and labels and axioms all match in this robot template. Don't parse the file, just look at it

id,rdfs label,editor preferred term,has curation status,alternative term,definition,definition
source,example of usage,term editor,editor note,curator note,ontology term requester,term tracker
item,logical type,parent,output

ID,A rdfs:label,A editor preferred term,AI has curation status,A alternative term SPLIT=|,A definition
SPLIT=|,A definition source SPLIT=|,A example of usage SPLIT=|,A term editor SPLIT=|,A editor note
SPLIT=|,A curator note SPLIT=|,A ontology term requester SPLIT=|,A term tracker item
SPLIT=|,CLASS_TYPE,C %,C has_specified_output some %

OBI:0003616,antibody purification of MHC class I protein complex,,,,,"An antibody purification of MHC
protein complex to separate a material entity into different compositions of which one fraction contains
a higher concentration of MHC class I protein complex, while others contain impurities and are not of
interest.",Randi Vita,,Randi VitalSebastian Duesing,,,,,equivalent,antibody purification of MHC protein

Review-based workflow - success!

Upon reevaluation, one inconsistency that stands out involves the mention of different species in the label versus the definition. For instance, the definition for the term with ID OBI:0003624 mentions a "rabbit MHC class I protein complex" in the definition, whereas the label and the output field reference a "pig MHC class I protein complex". This discrepancy indicates an error in either the label or the definition, as the species mentioned should be consistent across all fields to avoid confusion.

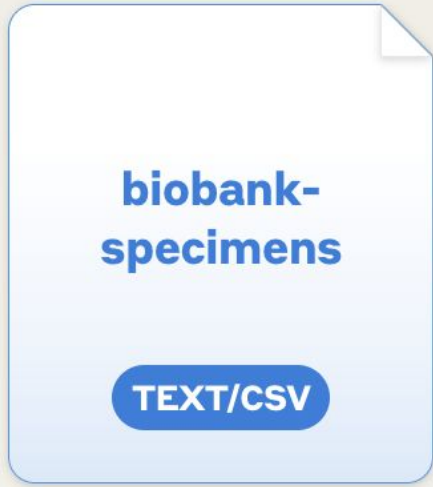
Review-based workflow: larger files

Example file: biobank specimens from OBI

- <https://gist.githubusercontent.com/cmungall/ec2a2cc2a453be40ab755bc1aa650520/raw/c8cbe6dec97ccbded52dbd9f6f9d0244d7daf9cb/biobank-specimens.csv>
- > 100 lines, many columns

Too large to paste into ChatGPT!

Alternative: claude.ai



Can you find any errors with definitions in the attached?

I found a few errors and inconsistencies in the definitions:

1. OBI:0002502 - The label and definition have a typo. It should be "cerebrospinal fluid specimen" instead of "cerbrospinal fluid specimen".
2. OBI:0002511 - The label and definition have a typo. It should be "vitreous humor specimen" instead of "vireous humor specimen".
3. OBI:0002613 - Typo in the definition. It should be "environmental material" instead of "environmenal material".
4. OBI:0002784 - The label states it is an "anterior nasal swab specimen" but the definition says it is derived from the "anterior nasal wall". The definition seems incorrect as a swab specimen should be collected from a surface, not derived from a structure.
5. OBI:2000016, OBI:2000017, OBI:2000018 - The logical type is inconsistent with the definition. The logical type says "collecting specimen with swab" but the definition indicates it is a specimen derived from a cell type, not collected with a swab.
6. Several terms like OBI:2000006, OBI:2000007 etc. have duplicate definitions under "definition" and "logical type" using slightly different wording.

Missing definition test

Example robot template with definitions missing:

- <https://gist.github.com/cmungall/2be046ef3f7eba16bb8a54b677280692>

This time let's use groq

| | A | B | C | D | E | F | G |
|---|-------------|------------------------------------|--------------------------------------------------------------|---|---|---|---|
| 6 | OBI:0002533 | colon specimen | A specimen that is derived from colon. | | | | |
| 7 | OBI:0002534 | lung specimen | A specimen that is derived from lung. | | | | |
| 8 | OBI:0002535 | esophagus muscularis mucosa specim | A specimen that is derived from esophagus muscularis mucosa. | | | | |
| 9 | OBI:0002536 | cerebral cortex specimen | A specimen that is derived from cerebral cortex. | | | | |
| 0 | OBI:0002537 | thyroid gland specimen | A specimen that is derived from thyroid gland. | | | | |
| 1 | OBI:0002538 | cerebellum specimen | A specimen that is derived from cerebellum. | | | | |
| 2 | OBI:0002539 | tib | A specimen that is derived from tibial nerve. | | | | |
| 3 | OBI:0002540 | coronary artery specimen | A specimen that is derived from coronary artery. | | | | |
| 4 | OBI:0002541 | spleen specimen | | | | | |
| 5 | OBI:0002542 | aorta specimen | | | | | |
| 6 | OBI:0002543 | atrial appendage specimen | | | | | |
| 7 | OBI:0002544 | esophagogastric junction specimen | | | | | |
| 8 | OBI:0002545 | ileum specimen | | | | | |
| 9 | OBI:0002546 | liver specimen | | | | | |
| 0 | OBI:0002547 | minor salivary gland specimen | | | | | |
| 1 | OBI:0002548 | omentum specimen | | | | | |
| 2 | OBI:0002549 | ovary specimen | | | | | |
| 3 | OBI:0002550 | sigmoid colon specimen | | | | | |
| 4 | OBI:0002551 | suprapubic skin specimen | | | | | |
| 5 | OBI:0002552 | testis specimen | | | | | |
| 6 | OBI:0002553 | uterus specimen | | | | | |
| 7 | OBI:0002554 | vagina specimen | | | | | |
| 8 | OBI:0002603 | areolar swab specimen | | | | | |
| 9 | OBI:0002604 | breast swab specimen | | | | | |
| 0 | OBI:0002605 | cheek swab specimen | | | | | |
| 1 | OBI:0002606 | nasopharyngeal swab specimen | | | | | |
| 2 | OBI:0002607 | oral swab specimen | | | | | |
| 3 | OBI:0002608 | oropharyngeal swab specimen | | | | | |



Here are the missing definitions for the given IDs:

1. OBI:0003482, lymph specimen, A specimen that is derived from lymphatic fluid.
2. OBI:0003483, digestive tract specimen, A specimen that is derived from the digestive tract.
3. OBI:0003595, nasal lavage fluid, A specimen collected through nasal cavity lavage that contains the reagents used to for the lavage process, organisms, cells, cellular secretions, and other biomaterials present in the nasal cavity space.
4. OBI:2000000, uterine cervix specimen, A specimen that is derived from the cervix of the uterus.
5. OBI:2000001, urethra specimen, A specimen that is derived from the

ROBOT Helper

ROBOT-Template Helper ▾



ROBOT-Template Helper

By Christopher Mungall 

Helps make robot templates to build ontologies

Make me an ontology of snacks

How do I make a definition with robot templates?

Make an ontology of cell types in the immune system

make an ontology of musical instruments

 Message ROBOT-Template Helper...



<https://bit.ly/robot-gpt>



You


Make an ontology of cell types in the immune system



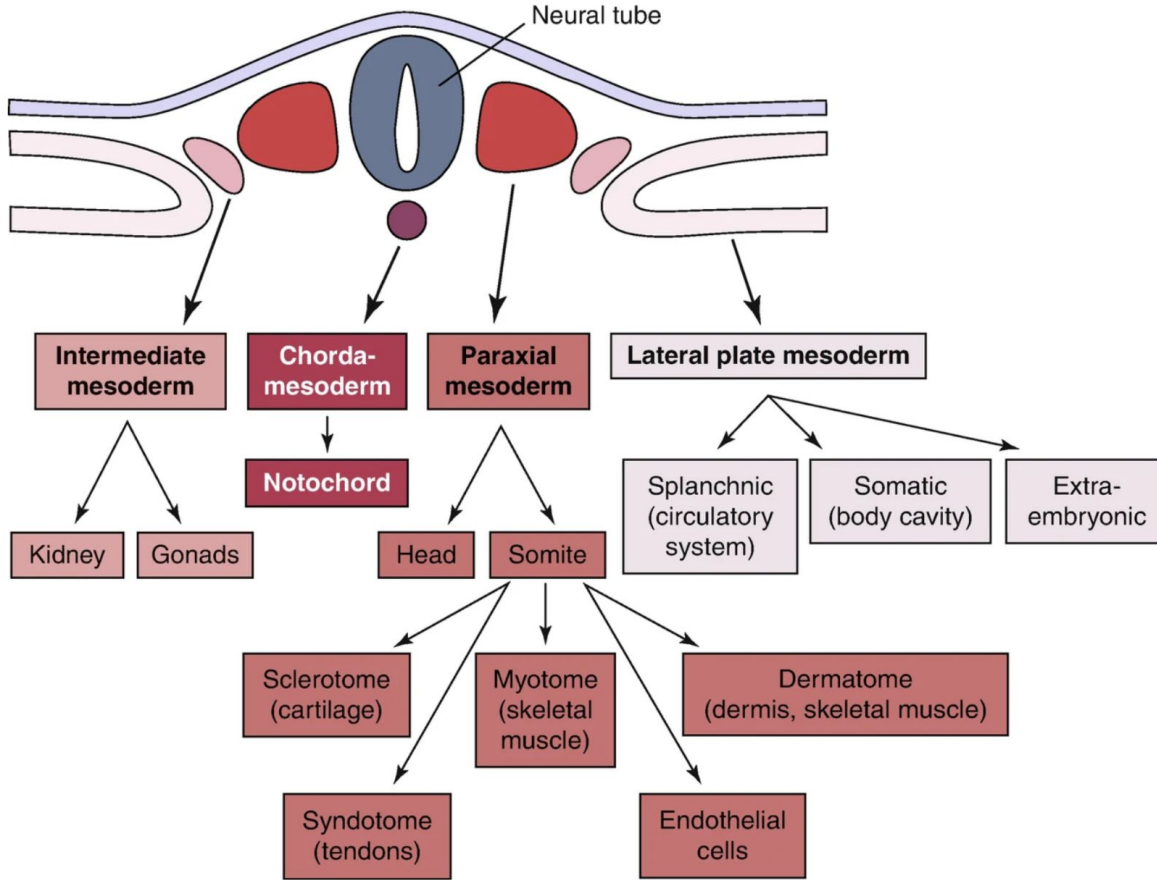
ROBOT-Template Helper

The uploaded document appears to be a template guide for converting tables to OWL format using ROBOT, rather than providing specific content about cell types in the immune system. Given this, I'll create a basic ontology of cell types in the immune system without relying on the uploaded document. This ontology will include a variety of cell types, with basic parent-child relationships and some example properties.

csv

 Copy code

```
identifier,name,parent,synonyms
ID,LABEL,SC %,A oboInOwl:hasExactSynonym
IMMUNE:1,Immune system,,Systema immunitatis
IMMUNE:2,Innate immune system,Immune system,
IMMUNE:3,Adaptive immune system,Immune system,
IMMUNE:4,Macrophage,Innate immune system,Monocyte-derived macrophage
IMMUNE:5,Neutrophil,Innate immune system,Polymorphonuclear leukocyte
IMMUNE:6,Dendritic cell,Innate immune system,
IMMUNE:7,T lymphocyte,Adaptive immune system,T cell
```



|

Here's the CSV format robot template:

CSV

 Copy code

```
identifier,name,parent,develops_into
ID,LABEL,SC %,A 'develops into'
EMBRYO:001,Neural tube,,Chorda-mesoderm|Paraxial mesoderm|Lateral plate mesoderm
EMBRYO:002,Intermediate mesoderm,EMBRYO:001,Kidney|Gonads
EMBRYO:003,Chorda-mesoderm,EMBRYO:001,Notocord
EMBRYO:004,Paraxial mesoderm,EMBRYO:001,Head|Somite
EMBRYO:005,Lateral plate mesoderm,EMBRYO:001,Splanchnic (circulatory system)|Somatic
EMBRYO:006,Notocord,EMBRYO:003,
EMBRYO:007,Head,EMBRYO:004,
EMBRYO:008,Somite,EMBRYO:004,Sclerotome (cartilage)|Myotome (skeletal muscle)|Dermato
EMBRYO:009,Splanchnic (circulatory system),EMBRYO:005,
EMBRYO:010,Somatic (body cavity),EMBRYO:005,
EMBRYO:011,Extra-embryonic,EMBRYO:005,
EMBRYO:012,Kidney,EMBRYO:002,
EMBRYO:013,Gonads,EMBRYO:002,
EMBRYO:014,Sclerotome (cartilage),EMBRYO:008,
```



You

what are the marker genes for Paraxial mesoderm?



Consensus



Talked to chat.consensus.app

The marker genes for Paraxial mesoderm have been investigated in various studies. Key genes involved in the specification of mesoderm to paraxial versus intermediate fates include:

1. **Foxc1 and Foxc2:** These forkhead genes regulate the establishment of paraxial versus intermediate mesoderm cell fates in the vertebrate embryo ([Wilm et al., 2004](#)).
2. **Wnt3a, Msgn1, and Tbx6:** These are key factors for paraxial mesoderm formation, with the absence of any one of these factors resulting in the complete absence of posterior somites ([Nowotschin et al., 2012](#)).
3. **pMesogenin1:** This bHLH class gene is specifically expressed in unsegmented paraxial mesoderm and its immediate progenitors, and it can drive nonmesodermal cells to assume a phenotype with molecular and cellular characteristics of early paraxial mesoderm ([Yoon et al., 2000](#)).

Questions?



DRAGON-AI evaluation

Chris Mungall

Lawrence Berkeley National Laboratory

OBO Academy

March 2024

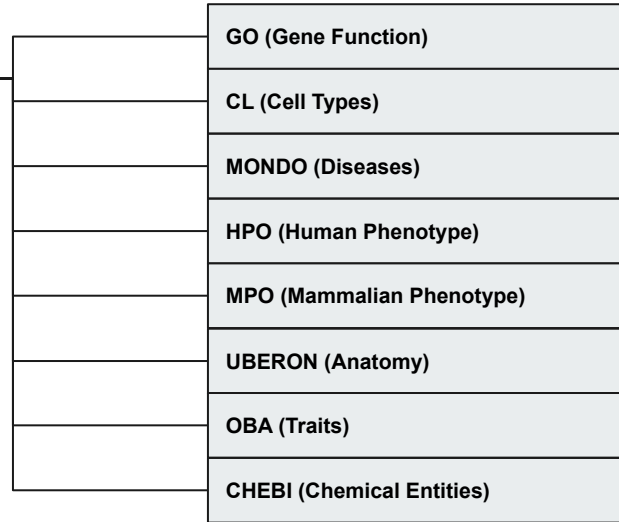
Ontologies are vital for Knowledge Bases



Ontologies: >1k

Terms: >14m

Ontologies (Examples)



<and many more>

DRKBs (Examples)



Ontologies and knowledge bases are living resources, continually updated and maintained

Can generative AI help maintain ontologies?

Dispelling Myths

“Manual curation is not scalable”

“Knowledge in the literature growing exponentially”

This is not true!

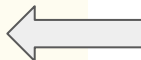
On expert curation and scalability: UniProtKB/Swiss-Prot as a case study

[Poux et al., *Bioinformatics* 2017](#)

Table 1.

Random sampling of PubMed

| | |
|-----------------------|-----------|
| Curatable | 19 |
| Not Priority | 17 |
| Not Curatable | |
| Out of scope | 452 |
| Redundant | 2 |
| High-throughput | 2 |
| Insufficient evidence | 5 |
| Review/comment | 3 |
| Total | 500 |



Can generative AI help maintain ontologies?

Dispelling Myths

“Manual curation is not scalable”

“Knowledge in the literature growing exponentially”

This is not true!

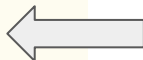
On expert curation and scalability: UniProtKB/Swiss-Prot as a case study

[Poux et al., *Bioinformatics* 2017](#)

Table 1.

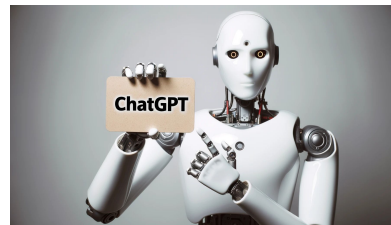
Random sampling of PubMed

| | |
|-----------------------|-----------|
| Curatable | 19 |
| Not Priority | 17 |
| Not Curatable | |
| Out of scope | 452 |
| Redundant | 2 |
| High-throughput | 2 |
| Insufficient evidence | 5 |
| Review/comment | 3 |
| Total | 500 |



Reality

- We could do with some help!
 - 1000s of open issues on GitHub trackers
- We should use **experts for high level tasks and automation for low level tasks**
- Particularly true for ontologies



Can AI help here?

Are LLMs alone the solution?



You

create a term in the plant trait ontology for the concept "Fine root C/leaf C ratio". Show in obo format



ChatGPT

To create a term in the Plant Trait Ontology (PTO) for the concept "Fine root C/leaf C ratio" in OBO (Open Biomedical Ontologies) format, we need to follow the standard conventions of OBO format. Here's an example of how this new term could be represented:

```
vbnet Copy code
[Term]
id: PT0:1234567
name: fine root C/leaf C ratio
namespace: plant_trait_ontology
def: "The ratio of carbon content in fine roots to the carbon content in l
synonym: "root to leaf carbon ratio" EXACT []
is_a: PT0:0000002 ! plant anatomical entity property
relationship: part_of PT0:0000203 ! root anatomy and morphology trait
relationship: part_of PT0:0000212 ! leaf anatomy and morphology trait
```



Challenge: LLMs hallucinate and have reliability issues



You

create a term in the plant trait ontology for the concept "Fine root C/leaf C ratio". Show in obo format

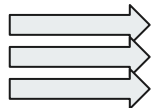


ChatGPT

To create a term in the Plant Trait Ontology (PTO) for the concept "Fine root C/leaf C ratio" in OBO (Open Biomedical Ontologies) format, we need to follow the standard conventions of OBO format. Here's an example of how this new term could be represented:

```
vbnet Copy code  
  
[Term]  
id: PT0:1234567  
name: fine root C/leaf C ratio  
namespace: plant_trait_ontology  
def: "The ratio of carbon content in fine roots to the carbon content in l  
synonym: "root to leaf carbon ratio" EXACT []  
is_a: PT0:0000002 ! plant anatomical entity property  
relationship: part_of PT0:0000203 ! root anatomy and morphology trait  
relationship: part_of PT0:0000212 ! leaf anatomy and morphology trait
```

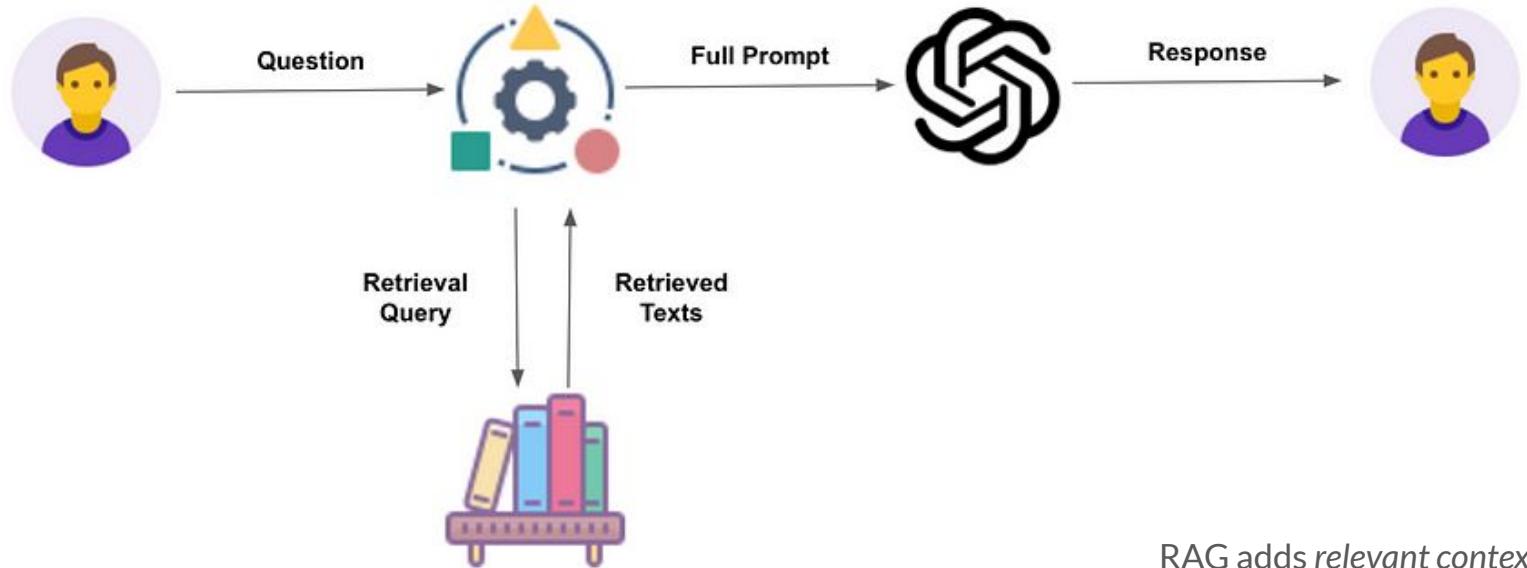
No such term



Too general a parent (ChatGPT doesn't know the ontology well enough to place a term)



Retrieval Augmented Generation (RAG) adds context



RAG adds *relevant context* to the prompt passed to the LLM

DRAGON-AI: Dynamic Retrieval Augmented Generation of Ontologies using AI

Create a new Class

Name spikelet floral color

IRI

New entity options...

Cancel OK

Annotations +

rdfs:label
spikelet floret color

definition
A spikelet floret morphology trait (TO:0000274) which is the color of a spikelet floret (PO:0009082).
database_cross_reference
Gramene:pankaj_jaiswal

Description: spikelet floret color

Equivalent To +
color
and ('characteristic of' some 'spikelet floret')

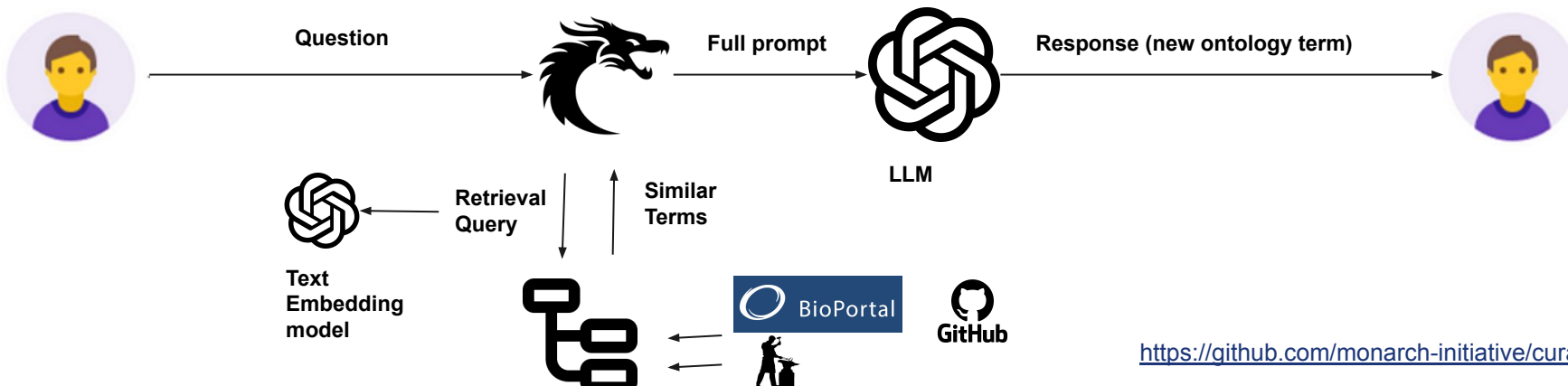
SubClass Of +
'flower color'
'spikelet floret morphology trait'

Definition (text)

Logical (OWL) definition

Relationships

Question
"Add term 'spiketet floral color'"



Challenge: How do we evaluate performance?

- **Subjectivity**
 - Ontology terms are constructs
 - *Multiple ground truths*
- **Heterogeneity**
 - Different knowledge domains present different challenges
 - *What works for one ontology or knowledgebase may not work for another*
- **Multifaceted**
 - Ontologies have a variety of *axiom* types (logical and textual)
 - *Not just link prediction*
- **Realistic scenarios**
 - Ontology development is collaborative
 - *How to emulate domain experts giving background knowledge*
- **Benchmarks**
 - How do we deal with test data leakage?
 - *All major LLMs have “seen” most ontology terms already*
- **Model proliferation**
 - Generalizability across models
 - *Different models are proliferating (Gemini, Mixtral, ...)*

Our evaluation approach

- **Subjectivity**
 - Ontology terms are constructs
 - *Manual evaluation using multiple experts*
- **Heterogeneity**
 - Different knowledge domains present different challenges
 - *Test on a variety of ontologies, evaluate on multiple criteria*
- **Multifaceted**
 - Ontologies have a variety of *axiom* types (logical and textual)
 - *Evaluate each axiom type separately*
- **Realistic scenarios**
 - Ontology development is collaborative
 - *Evaluate realistic workflows, use of GitHub issues as background knowledge*
- **Benchmarks**
 - How do we deal with test data leakage?
 - *Test using only terms from after model cutoff*
- **Model proliferation**
 - Generalizability across models
 - *Test open and closed models*

Evaluating AI-Generated Definitions against Human Curated Definitions

24 Evaluators volunteered

- editors, curators, and users

Evaluators considered terms generated by different AI models as well as manually authored ones

- Evaluators worked independently
- Evaluators were blinded to source of term

Scored using different criteria (1-5)

- Accuracy
- Consistency
- Overall score

Evaluators also indicated confidence (1-5)



<https://github.com/monarch-initiative/dragon-ai-results>



tair



Evaluation on 10 ontologies

| Ontology | Name | Typical Usage | Terms Tested | Tasks Evaluated |
|----------|---------------------------------------|--------------------------------------------------------|--------------|-----------------|
| CL | Cell type ontology | <i>Single cell annotation</i> | 50 | R,D,L |
| ENVO | Environment Ontology | <i>Microbiome annotation</i> | 50 | R,D,L |
| FOODON | Food Ontology | <i>Nutrition, food safety</i> | 50 | R,D,L |
| GO | Gene Ontology | <i>Gene function analysis</i> | 50 | R,D,L |
| HP | Human Phenotype Ontology | <i>Disease/gene annotation, variant prioritization</i> | 50 | R,D |
| MONDO | Mondo Disease Ontology | <i>Disease classification</i> | 50 | R,D,L |
| MP | Mammalian Phenotype Ontology | <i>Gene/genotype annotation</i> | 50 | R,D |
| OBA | Ontology of Biological Attributes | <i>GWAS, eQTL</i> | 50 | R,D,L |
| OBI | Ontology of Biomedical Investigations | <i>Experimental methods</i> | 50 | R |
| UBERON | Anatomy Ontology | <i>Gene expression</i> | 40 | R,D,L |

Inclusion criteria

- Pass OBO Dashboard
- Impactful for NIH DRKBs
- Actively developed
- Cross-section of domains
- Diversity of styles of ontology development

Tasks

- Relationships
 - Classification
 - All relationships
- Definitions (text)
- Logical axiomatization

Models and configurations tested



Gpt-3.5-turbo (OpenAI)

Closed, inexpensive, API only



Gpt-4 (OpenAI)

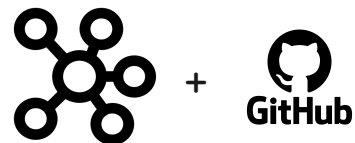
Closed, expensive, API only



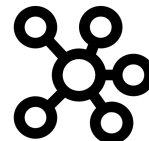
nous-hermes-13b

Open, API or download and run locally (huggingface.co)

X



RAG over ontology + GitHub issues



RAG over ontology only

Results: Humans beat AI!



Terms authored by humans scored **better** than those authored by AI



Relationship Prediction Task

| method | model | SubClassOf Task | | | All Relationship Types Task | | |
|---------------|-----------------|-----------------|-------------|--------------|-----------------------------|--------------|-------------|
| | | precision | recall | f1 | precision | recall | f1 |
| DRAGON | gpt-3.5-turbo | 0.831 | 0.352 | 0.494 | 0.746 | 0.392 | 0.514 |
| DRAGON | gpt-4 | 0.889 | 0.44 | 0.588 | 0.797 | 0.456 | 0.58 |
| DRAGON | nous-hermes-13b | 0.68 | 0.273 | 0.39 | 0.597 | 0.292 | 0.392 |
| OWL Reasoner | n/a | 1.0* | 0.337 | 0.504 | n/a | n/a | n/a |

(results aggregated across ontologies)

- Relationships were largely correct
- Precision scores may undercount due to missing relationships in ontology



Definition evaluation: Human curated definitions (narrowly) beat AI generated ones

| method | model name | Accuracy (1-5) | Score (1-5) | Consistency (1-5) |
|----------------|-----------------|----------------|-------------|-------------------|
| DRAGON | gpt-3.5-turbo | 4.018 | 3.598 | 3.7 |
| DRAGON | gpt-4 | 3.945 | 3.543 | 3.664 |
| DRAGON | nous-hermes-13b | 3.741 | 3.361 | 3.548 |
| curator | human | 4.332 | 4.06 | 4.107 |

} 2nd best (joint)

best

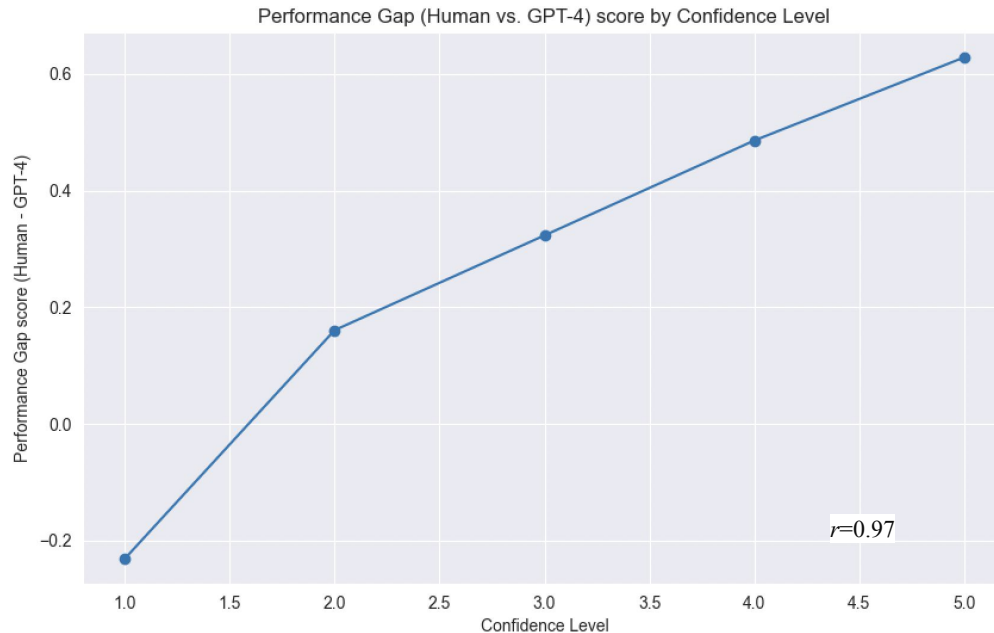
gpt-4 vs gpt-3.5 not statistically significant

Closed models vs open model: **statistically significant**

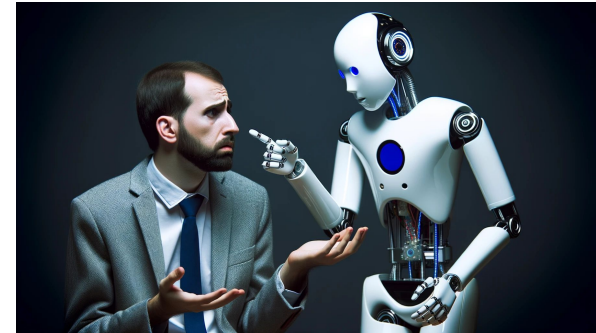
Humans vs AI: **statistically significant**

Adding GitHub issues as a background source boosted performance by **0.2** on average (only tested on subset)

Note of caution: Less-confident evaluators gave AI better scores



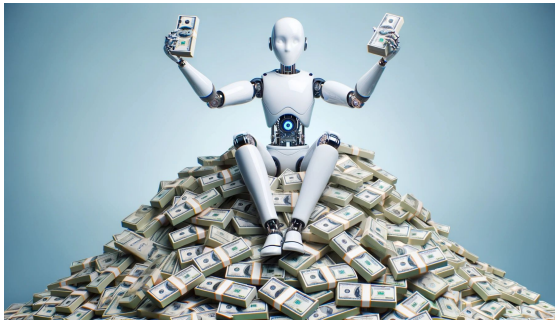
“Gaslight” effect



Lessons from software engineering?

June 2023

“Estimated GDP boost of **\$1.5trillion** by 2030”



Domhke et al (2023) *Sea Change in Software Development: Economic and Productivity Analysis of the AI-Powered Developer Lifecycle* [10.48550/arXiv.2306.15033](https://arxiv.org/abs/10.48550/arXiv.2306.15033)

Lessons from software engineering?

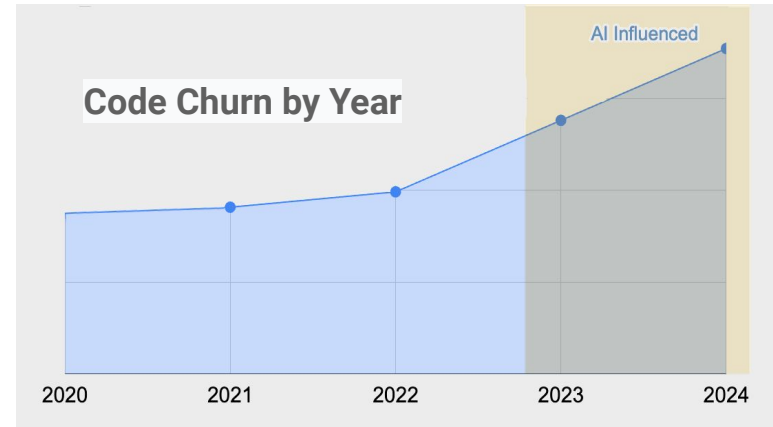
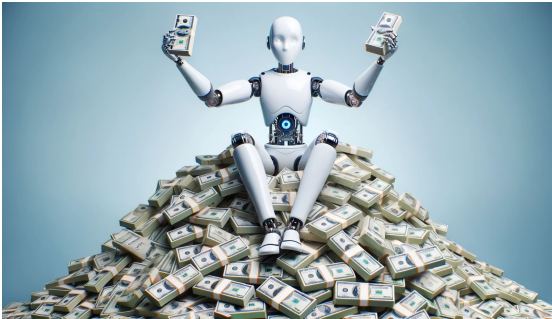
June 2023



Jan 2024

“Estimated GDP boost of **\$1.5trillion** by 2030”

2023 Data Shows Downward Pressure on Code Quality ⇒ **Technical Debt**



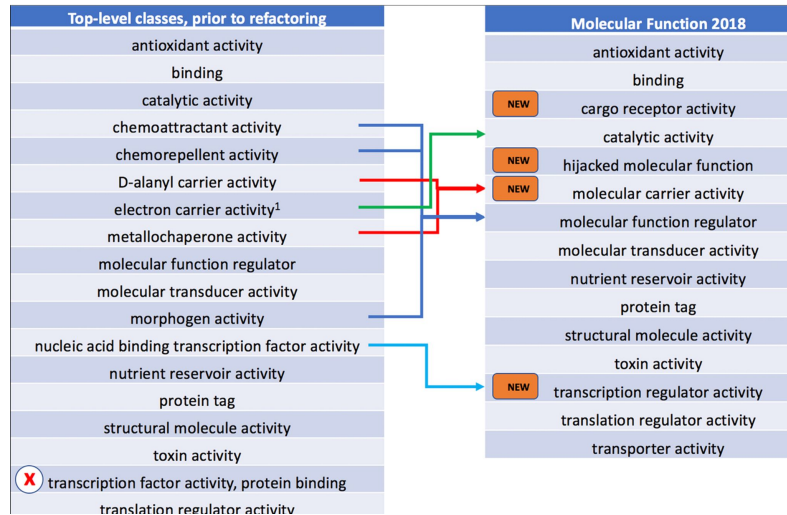
Domhke et al (2023) *Sea Change in Software Development: Economic and Productivity Analysis of the AI-Powered Developer Lifecycle* [10.48550/arXiv.2306.15033](https://arxiv.org/abs/10.48550/arXiv.2306.15033)

Harding W, Kloster M (2024) *Coding on Copilot (GitClear analysis of 150m lines of code)*
https://www.gitclear.com/coding_on_copilot_data_shows_ais_downward_pressure_on_code_quality

Limitations of our study

- Test set was restricted to recently added terms to mature ontologies which may not be representative
- Most ontology work is **refactoring**, *not simply accumulating terms*

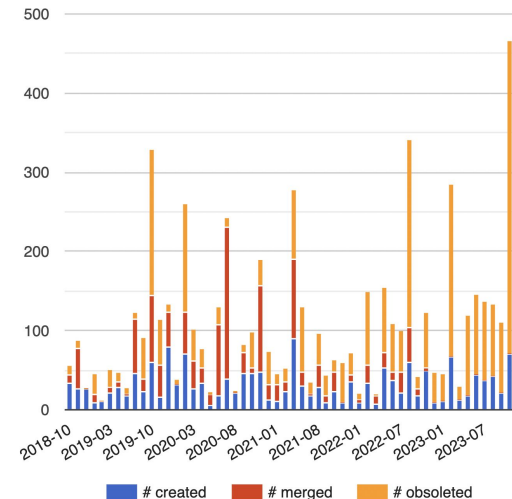
Ontology Refactoring



GO consortium (2019), *Nucleic Acids Research*

Obsolescence and merging

Changes in GO terms between releases



geneontology.org/stats.html

DRAGON-AI part of Monarch suite of LLM tools



1
Derive structured knowledge from unstructured sources

SPIRES / OntoGPT

Caufield et al (2024) *Bioinformatics*, btae104



2
Interpret and summarize omics data and lists of entities

TALISMAN

Joachimiak et al (2023) *arxiv*, 2305.13338



3
Enhance GPT with trusted information from key knowledge providers

Phenomics Assistant

O'Neil et al (2023) *biorxiv*, 2024.01.31.578275

4
Find evidence for statements and hypotheses

CurateGPT

github.com/monarch-initiative/curate-gpt



5
Constructing Ontologies and KGs

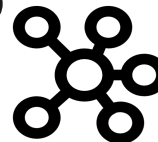
DRAGON-AI

Toro et al (2023) *arxiv*, 2312.10904

6
Integrate different standards and vocabularies

MapperGPT

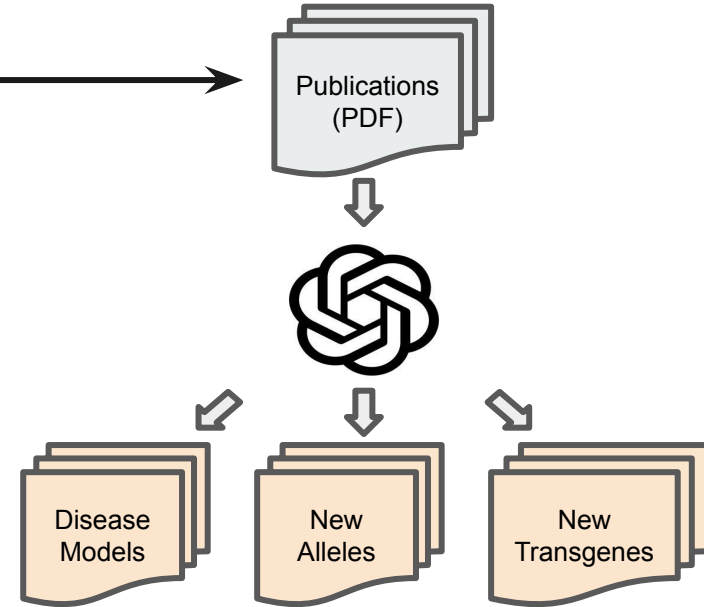
Matentzoglou et al (2023) *arxiv*, 2310.03666



Alliance of Genome Resources AI curation tools

- Exploring models with large context windows for full papers
- LLM-based **virtual biocurator assistant** to help curators extract complex information from papers
- Automate article **“trialoging”** using AI
- **Document classification** using AI
- Improving **gene summaries** using AI/LLM
- Streamline AI/ML pipelines creation, evaluation, comparison and versioning through a **centralized management framework**

```
{  
  
  "reasoning": "The publication describes the L138F mutation in human Orai1, which can cause severe myopathy, and its corresponding mutation L210F in dOrai of Drosophila melanogaster. Molecular dynamics simulations were performed for the WT and the L210F mutant channels to understand how the L210F mutation activates the channel. This indicates that Drosophila is used as a model to study the activation mechanism of a mutation relevant to a human disease, which fits the criteria for Drosophila models of human disease as described in the prompt."  
  
  "triage_result": true,  
  
}
```



Conclusions

- High quality **knowledge bases and ontologies** can be used to improve accuracy of LLMs using Retrieval Augmented Generation (RAG)
 - RAG works best with *existing high quality curated knowledge*
- The combination has the potential to be useful **assistants** to ontology developers and knowledge base curators
 - Results **should be vetted by experts**
 - Care should be taken not to accumulate **semantic debt**
- Integration into **existing workflows and tools** is essential
 - Ontology Development Environments (e.g. Protege)

Acknowledgements

Sabrina Toro

Anna V Anagnostopoulos

Sue Bello

Kai Blumberg

Rhiannon Cameron

Leigh Carmody

Harry Caufield

Alexander D Diehl

Damion Dooley

William Duncan

Petra Fey

Pascale Gaudet

Nomi L Harris

Marcin Joachimiak

Leila Kiani

Tiago Lubiana

Monica C Munoz-Torres

Shawn O'Neil

David Osumi-Sutherland

Aleix Puig

Justin P Reese

Leonore Reiser

Sofia Robb

Troy Ruemping

James Seager

Eric Sid

Ray Stefancsik

Magalie Weber

Valerie Wood

Peter Robinson

Melissa A Haendel

NHGRI

HG010860 **Phenomics First**

HG012212 **Gene Ontology**

U24HG010859 **Alliance**

NIH OD

R24 OD011883 **Monarch Initiative**

NSF [OAC-2112606 **ICICLE**]

DOE [DE-AC0205CH11231]