“Standing on the Shoulders of Giants” is Becoming Harder

- Dozens of new models in publications each year. Difficult to:
  - Remain well-informed as a researcher.
  - Educate the general public.
  - Constructively build upon existing work.
- Models are not like long-lasting software products:
  - Partially described by papers, results and other artifacts.
  - Code might not even exist.
  - Definitely not implemented with extensibility in mind.
- Early days of the COVID-19 pandemic: top-level policy decisions based on old undocumented code.

**DARPA’s Automating Scientific Knowledge Extraction and Modeling (ASKEM) project [1]**
- Develop tools will enable experts to maintain, reuse, and adapt large collections of heterogeneous data, knowledge and models.

**Entity Annotation is a Basic Building Block**

**Code Self-Documentation is Often Lacking**
- In order to reuse/extend a model, scientists must understand it.
- Annotations in the code itself might be insufficient.
- But the knowledge exists in the original model description.
- **Task 1**: Can we annotate code elements with their descriptions from text/equations?

**Different Papers, Different Terms**
- Terminology might not be standardized across works, making model comparison harder.
- **Task 2**: Can we map model terms to a single source, like a Domain Knowledge Graph (DKG)?

**Models Without Data are Unusable**
- To evaluate a model, data must be provided for each variable.
- The data schema might not match the variable definitions.
- **Task 3**: Can we find the most appropriate data for each variable?

**Large Language Models to the Rescue!**

**Close Enough for Humans? Close Enough for GPT-3 [4]!**
- Even though terminology for the same variable might differ across sources, the terms used are usually semantically similar enough for a human.
- Models like GPT-3 are also able to pick up on this similarity!
- After appropriate prompt engineering, we can use GPT-3 for the three tasks above.

**Giving Downstream Users an Editable Model**
- Our API can extract a graph description of a model from code.
- It can then annotate each variable with text descriptions, equations, datasets and/or DKG terms, whenever available.
- Downstream ASKEM teams can then visualize this model.
- Users can leverage the associated annotations to understand the model and evolve it as needed.

---

**Notes**