# Reasoning over ontologies using Large Language Models

*This manuscript (permalink) was automatically generated from cmungall/gpt-reasoning-manuscript@fba1ca3 on May 30, 2023.*

## Authors


- **John Doe**
  XXXX-XXXX-XXXX-XXXX · johndoe · johndoe · @johndoe@mastodon.social
  Department of Something, University of Whatever · Funded by Grant XXXXXXXX

- **Chris Mungall** ✉
  0000-0002-6601-2165 · cmungall
  Environmental Genomics and Systems Biology Division, Lawrence Berkeley National Laboratory, Berkeley, CA, 94720

✉ — Correspondence possible via GitHub Issues or email to Chris Mungall <cjmungall@lbl.gov>.

# Abstract

Reasoning is a core component of human intelligence, and a key goal of AI research. Reasoning has traditionally been the domain of symbolic AI, but recent advances in deep learning and in particular Large Language Models (LLMs) such as GPT-3 seem to suggest that LLMs have some latent reasoning ability.

To investigate this, we created a GPT-based reasoning agent that is intended to perform ontological reason using a few-shot learning approach, using instruction prompting and in-context examples. We also created a series of benchmarks to test ontological reasoning ability in LLMs and other systems.

Our results indicate that GPT is a poor reasoner, and is only able to perform ontological reasoning on some of the simplest tasks. Even on these simple tasks, results are highly variable, with performance degrading as the size of the ontology and the complexity of the explanation increases. In the cases where it does successfully perform the task, this seems to essentially be an advanced pattern-based form of lookup.

Our results indicate that a maximalist approach to using LLMs may be limiting, and that to be successful AI should use hybrid strategies.


# Introduction

## Motivation

The field of Artificial Intelligence (AI) has historically been dichotomized into two schools of thought: symbolic AI and connectionist AI. Symbolic AI is exemplified by ontologies and knowledge bases which represent knowledge as a set of symbols together with rules for manipulating those symbols. Connectionist AI is exemplified by deep learning and in particular Large Language Models (LLMs).

Recent advances in LLMs have shown exceptional abilities in tasks such as question-answering and text summarization, with the use of in-context learning able to substitute for task-specific training[1]. The abilities of the latest generation of LLMs such as GPT-4 seem to suggest that LLMs may be able to perform reasoning tasks that were previously the domain of symbolic AI, in the same way that a human may incorporate reasoning as a part of natural language question-answering. The implication here is that symbolic AI and classical deductive reasoning, if this is an emergent ability of LLMs. However, the ability of LLMs to reason over knowledge bases has never been systematically evaluated.

## Ontological Reasoning

Ontology reasoning is a form of reasoning that is based on the structure of a knowledge base or ontology.

Ontology reasoning underpins... TODO

## OWL Benchmark Datasets

TODO

- OWL2Bench
- LUBM

## Natural Language Benchmark Datasets

Existing datasets for testing reasoning ability include LogiQA [2] and ReClor [3]. These are aimed at testing reasoning abilities in the context of natural language understanding. For example, LogiQA contains as instance with paragraph "*David knows Mr. Zhang's friend Jack, and Jack knows David's friend Ms. Lin. Everyone of them who knows Jack has a master's degree, and everyone of them who knows Ms. Lin is from Shanghai.*" and question "*Who is from Shanghai and has a masters degree?*". Liu et al adapted these for evaluation using prompt-based LLMs such as GPT [4]..

## Contributions

We make the following contributions:

- We have curated a collection of benchmarks for testing ontological reasoning ability
- We created a GPT-based reasoning agent that is intended to perform ontological reason using a few-shot learning approach, using instruction prompting and in-context examples.
- We have evaluated the reasoning ability of GPT-3.5-turbo and GPT-4 on these benchmarks

# Methods

TEST: 1

## Semi-automatic generation of reasoning benchmarks from ontologies

We created a methods to generate 6 classes of benchmarks

| code | task | description |
|---|---|---|
| sat | OntologyCoherencyTask | A task to determine if an ontology is coherent. There should be a single answer, which is a boolean. |
| indirect | EntailedIndirectSuperClassTask | A task to determine the indirect superclasses of a class. |
| superc | EntailedTransitiveSuperClassTask | A task to determine the all transitive superclasses of a class. |
| expr | EntailedSubClassOfExpressionTask | A task to determine the subclasses of a class expression. |
| dir-sup | EntailedDirectSuperClassTask | A task to determine the direct superclasses of a class. Includes those entailed by other axioms, e.g. equivalence axioms. Context: a standard pattern in bio-ontologies is to infer the structure of one ontology from another - e.g. the metabolic process branch in GO may be entailed by GO equivalence axioms plus the IS_A links in CHEBI. |
| mrca | MostRecentCommonSubsumerTask | A task to determine the most specific common ancestors. |

| code | task | description |
|---|---|---|
| abox | ABoxTask | A task to infer assertions over property chains and transitvity in aboxes. |

➜

## GPT-based reasoning agent

We created a GPT-based reasoning agent that is intended to perform ontological reason using a few-shot learning approach, using instruction prompting and in-context examples.

The reasoning agent is implemented as part of the OntoGPT system.

3 methods for performing reasoning were implemented:

- direct
- post-hoc explanation-based
- chain-of-thought reasoning

### Direct reasoning

# Results

## Core Results

| | task | abox | expr | indirect | mrca | sat | superc |
|---|---|---|---|---|---|---|---|
| **model** | **mthd** | | | | | | |
| **gpt-3.5-turbo** | **basic** | 0.642 | 0.718 | 0.739 | 0.067 | **0.000** | 0.805 |
| | **cot** | 0.305 | 0.623 | 0.587 | 0.100 | **0.000** | 0.584 |
| | **expl** | 0.571 | 0.756 | 0.562 | 0.033 | **0.000** | 0.689 |
| **gpt-4** | **basic** | **1.000** | 0.844 | **0.940** | 0.150 | **0.000** | **0.961** |
| | **cot** | 0.929 | 0.664 | 0.814 | 0.150 | **0.000** | 0.794 |
| | **expl** | 0.994 | **0.849** | 0.813 | **0.183** | **0.000** | 0.960 |

## Obfuscation

| | task | abox | expr | indirect | mrca | sat | superc |
|---|---|---|---|---|---|---|---|
| **model** | **method** | | | | | | |
| **gpt-3.5-turbo** | **basic** | 0.000 | 0.270 | 0.256 | 0.000 | **0.000** | 0.426 |
| | **chain_of_thought** | 0.000 | 0.131 | 0.161 | 0.100 | **0.000** | 0.390 |
| | **explanation** | 0.000 | 0.216 | 0.207 | 0.050 | **0.000** | 0.471 |
| **gpt-4** | **basic** | **0.764** | 0.551 | **0.837** | **0.200** | **0.000** | **0.912** |
| | **chain_of_thought** | 0.568 | **0.818** | 0.506 | **0.200** | **0.000** | 0.718 |

| | task | abox | expr | indirect | mrca | sat | superc |
|---|---|---|---|---|---|---|---|
| **model** | **method** | | | | | | |
| | **explanation** | 0.631 | 0.728 | 0.701 | 0.167 | **0.000** | 0.812 |

### Effect of chain lengths

## Discussion

blah

## Conclusions

blah

# References


1. **Language Models are Few-Shot Learners**
   Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, … Dario Amodei
   *arXiv* (2020) https://doi.org/gpmv43
   DOI: 10.48550/arxiv.2005.14165

2. **LogiQA: A Challenge Dataset for Machine Reading Comprehension with Logical Reasoning**
   Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, Yue Zhang
   *arXiv* (2020) https://doi.org/gr9cn5
   DOI: 10.48550/arxiv.2007.08124

3. **ReClor: A Reading Comprehension Dataset Requiring Logical Reasoning**
   Weihao Yu, Zihang Jiang, Yanfei Dong, Jiashi Feng
   *arXiv* (2020) https://doi.org/gr9cn4
   DOI: 10.48550/arxiv.2002.04326

4. **Evaluating the Logical Reasoning Ability of ChatGPT and GPT-4**
   Hanmeng Liu, Ruoxi Ning, Zhiyang Teng, Jian Liu, Qiji Zhou, Yue Zhang
   *arXiv* (2023) https://doi.org/gr9cn6
   DOI: 10.48550/arxiv.2304.03439