From Models to Microtheories: Distilling a Model's Topical Knowledge for Grounded Question Answering

Nathaniel Weir, Bhavana Dalvi Mishra, Orion Weller, Oyvind Tafjord,
Sam Hornstein, Alexander Sabol, Peter Jansen, Benjamin Van Durme, Peter Clark

Ai2 · ICLR International Conference On Learning Representations · Johns Hopkins University

Motivation

Chain-of-thought materializes knowledge for **one** question. What about a whole **topic** of questions?

Can we create a method for extracting and reasoning over a discrete representation of an LLM's beliefs?

E.g. what does GPT-4 know about mechanical physics?

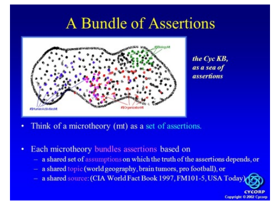
Inspiration

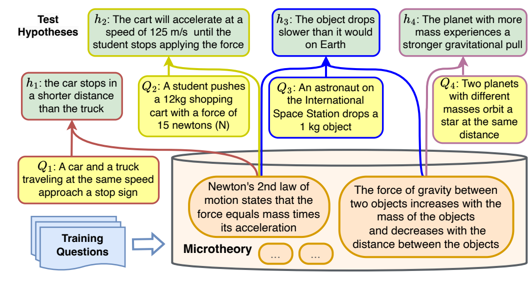
Humans develop and teach each other general ***microtheories*** about a topic space
e.g. you learned the "magnetism" microtheory from your school teacher

A microtheory serves as **representational buffer** that allows us to operationalize our understanding of an area. Can we apply the same idea to neural reasoning algorithms?

We define a **microtheory** as a concise, generalizable set of core facts about a domain

The **Cyc project** used a Similar idea in the 90s/00s

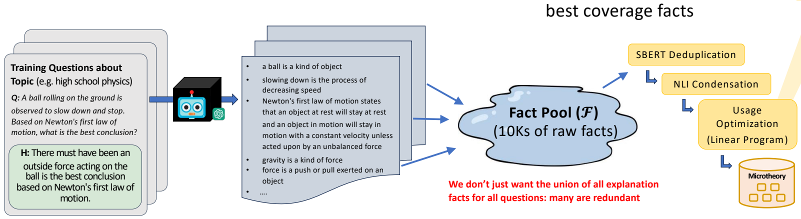




Check out the paper!

Construction Pipeline

Step 1: Extract Large, Redundant Fact Pool

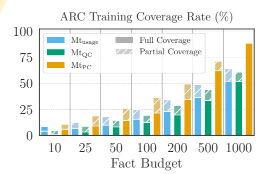


Usage Optimization

Step 2: Distill out semantic redundancies and find n best coverage facts

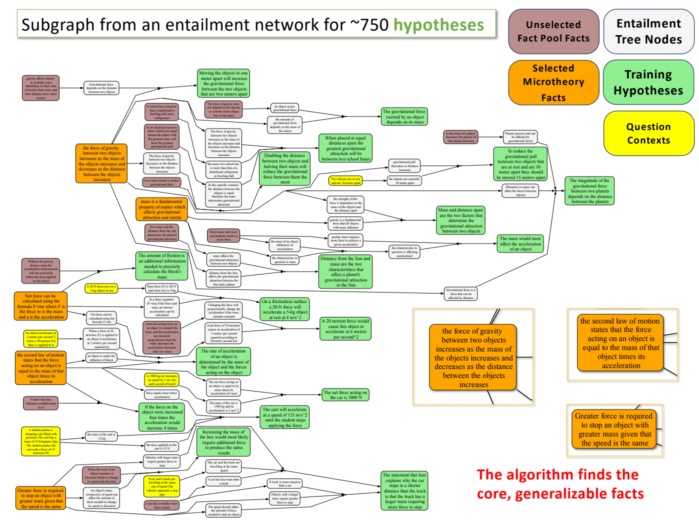
1. Using GPT-4 entailment engine, find all combinations of facts from condensed fact pool $\mathcal{C}$ that are an "argumentative basis" for each train hypothesis
$$L_i = [\text{fact}_1, \text{fact}_2, \text{fact}_3, \dots]$$
2. Linear program finds optimal n facts for max partial coverage (PC) of a vector for each hypothesis
(We also tried full question coverage (QC) and sorting by question usage frequency, those work less well)

$$n\text{-}\mathbf{Mt}_{\mathrm{PC}} = \operatorname*{arg\,max}_{\mathcal{M}\subseteq\mathcal{C},|\mathcal{M}|=n} \sum_{j=1}^{d} \operatorname*{arg\,max}_{L_{kj}\in\textsc{LeafSets}(\textsc{engine}(h_j,\mathcal{C};Q_j))} \frac{|\mathcal{M}\cap L_{kj}|}{|L_{kj}|}$$



	ARC	MedQA
Dataset Details		
# Questions	854/127/249	641/94/186
# Topic Clusters	9	4
Ext. Corpus	Wikipedia	Wiki+Textbooks
Training Extraction Results		
$\lvert\mathcal{F}\rvert$	20,561	16,722
$\lvert\mathcal{C}\rvert$	8,722	10,908
# Qs with Proof	744 (89%)	562 (88%)
Min #Fs to Cover	800	890
Fact/Q Ratio	1.08	1.58

Subgraph from an entailment network for ~750 hypotheses

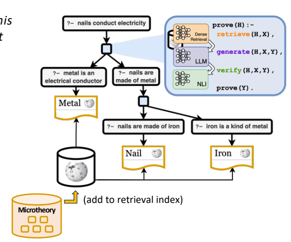


The algorithm finds the core, generalizable facts

Application to Entailment Engine

Does this hypothesis follow from this knowledge corpus? (and with what confidence?)

- Hooks up to any text retrieval corpus
- Uses LLMs/classifiers for generating + verifying decompositions
- Accepts a hypothesis if it can fully ground it to the corpus
  - Identifies different **"argumentative bases"** for supporting hypothesis
- As you give it more (useful) facts, it should reason better



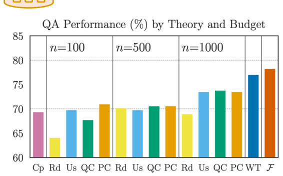


Figure 7: Question Answering performance on ARC topical test questions using Wikipedia plus various Mts (**Random**, **Usage**, **QC**, **PC**) as knowledge sources.

Engine QA Evaluation

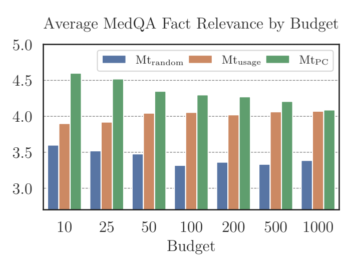
Adding 1000-fact Mts improves QA w/Engine by 4%

Adding full fact pool ($\mathcal{F}$) improves QA by 9%

Relevance Evaluation

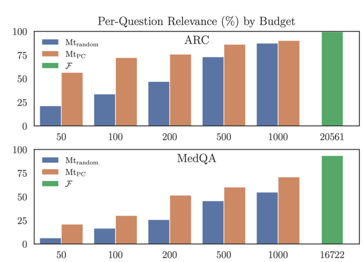
Human:
We recruited senior medical students to annotate relevance of MedQA microtheory facts

Microtheory optimized for partial coverage (PC) has higher scores than random selection or sorting by usage frequency



Automatic:
We asked GPT-4 whether *at least one* fact in a microtheory is useful for answering a test question correctly

PC-optimized microtheory has substantially higher relevance rate than random



Takeaways

We present a method for materializing a model's latent, topical knowledge into a discrete **microtheory** articulating the model's core, reusable knowledge about the topic

We explore ways to identify the most core n facts to explain training hypotheses from a large knowledge pool

Microtheories can improve a entailment engine's ability to ground hypotheses and correctly answer questions