Can LMs Learn New Entities from Descriptions? Challenges in Propagating Injected Knowledge

Yasumasa Onoe, Michael J.Q. Zhang, Shankar Padmanabhan, Greg Durrett, Eunsol Choi

TEXAS The University of Texas at Austin

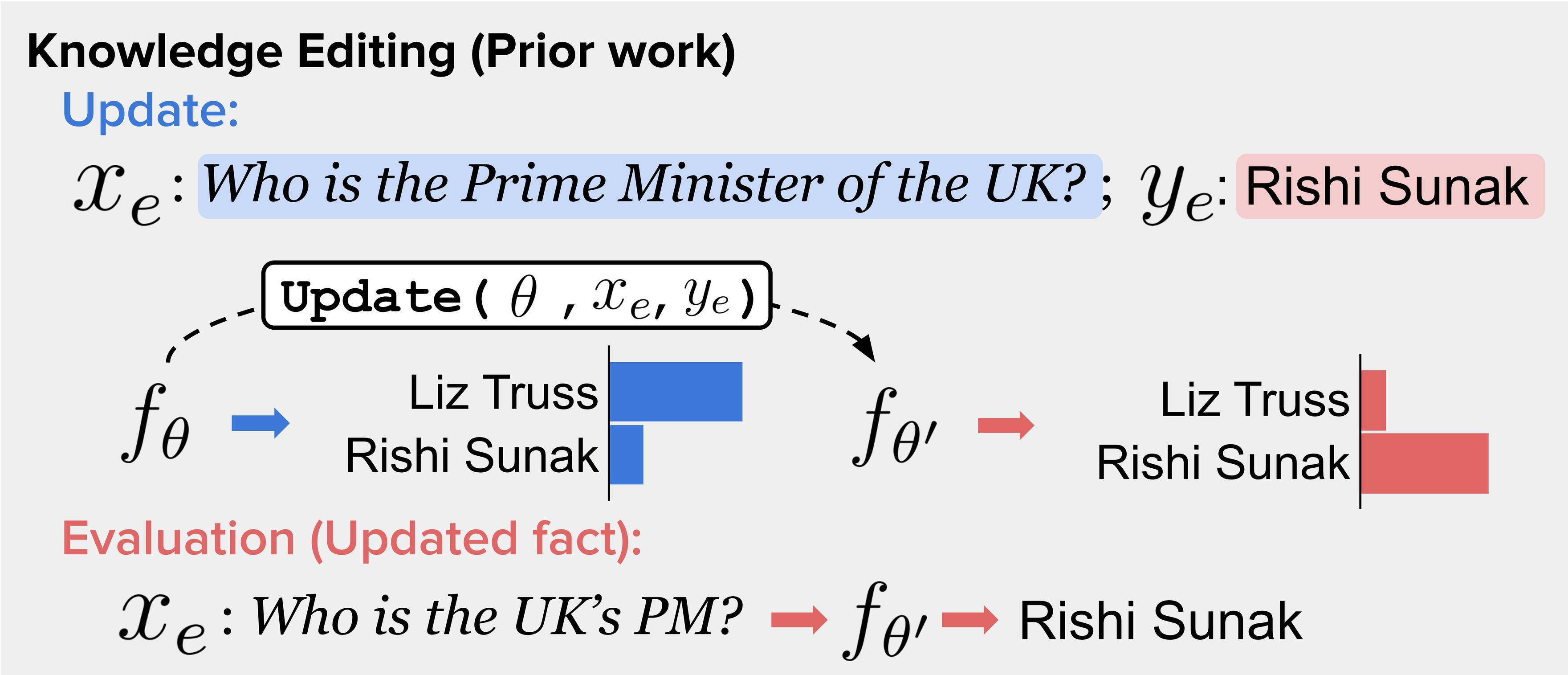
TAUR LAB

Motivation

Prior work has investigated knowledge editing in pre-trained LMs, updating model parameters to alter outputs to match what users want. We focus specifically on injecting **new entities** into models.

RQ1: Can LMs make inferences based on updated knowledge?
➡ **We propose a new task called Entity Knowledge Propagation (EKP).**

RQ2: How do SOTA knowledge editing methods perform on EKP?
➡ **We compare fine-tuning, MEND, ROME, and in-context use of the definition on two datasets.**



Entity Knowledge Propagation: when we teach an LM a new entity, can the model make inferences about it?

We **update** an LM on a definition sentence of a **new entity** using any KE method such as finetuning, MEND, or ROME.

Update:

d_e: ***Rishi Sunak** is a British politician who has served as Prime Minister of the United Kingdom.*

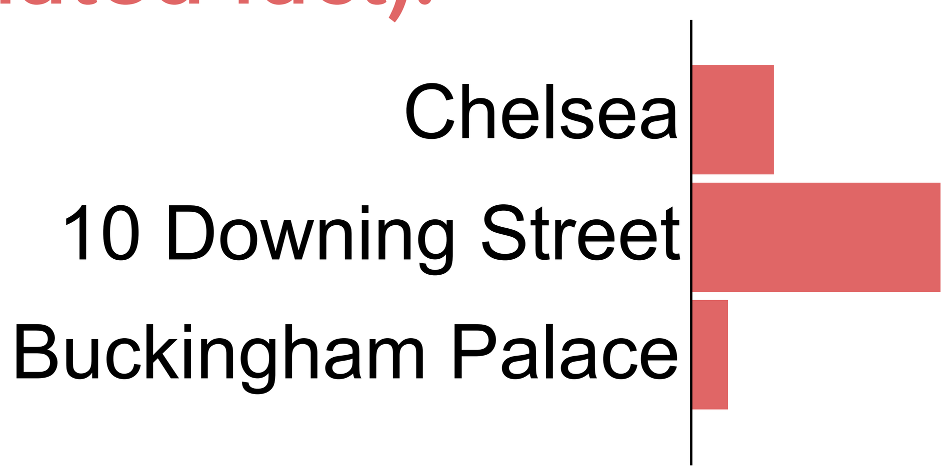
f_θ - - - $\boxed{\texttt{Update(}\ \theta\ ,d_e\texttt{)}}$ - - ➤ $f_{\theta'}$

Base LM — Updated LM

The updated LM is evaluated on a probe sentence. This could be a cloze-style task such as ECBD.

Evaluation (Inference based on the updated fact):

x_e : *Rishi Sunak lives at [MASK].* ➡ $f_{\theta'}$ (Updated LM) ➡



Experiments

Datasets

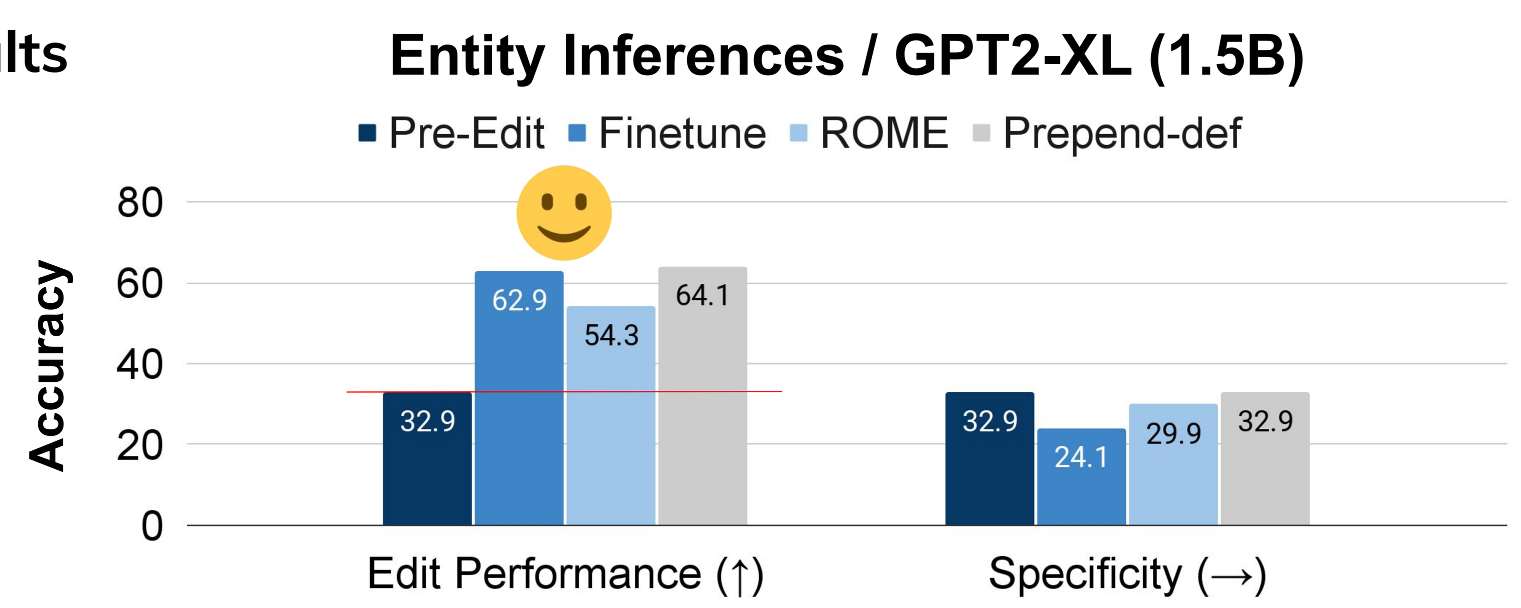
1. Entity Inferences (**new in this work**)
   - Manually crafted probe sentences using templates

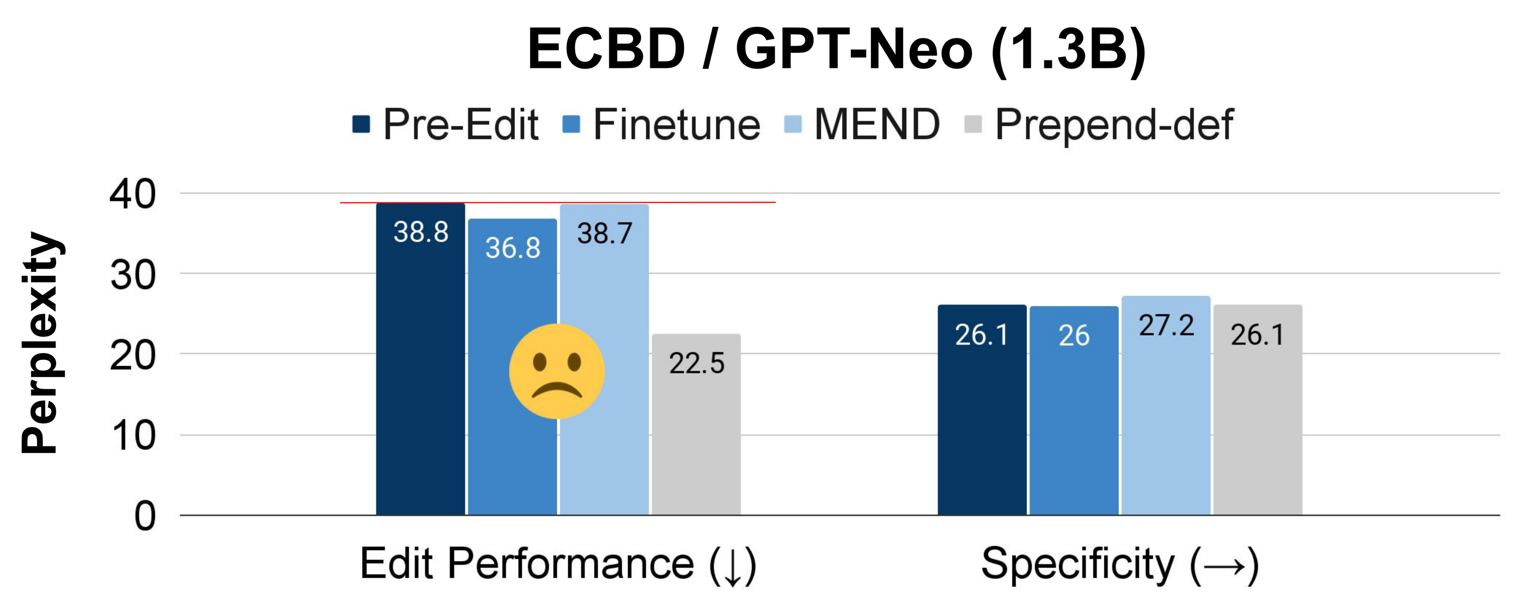
Definition: Hurricane Nana was a minimal Category 1 hurricane that caused moderate damage across Belize in early September 2020.
Sentence: Hurricane Nana (2020) totally [MASK] my house.
Entity: Hurricane Nana
Options: acted, brewed, built, destroyed,...
Label: destroyed

2. Entity Cloze By Date (ECBD, Onoe et al., 2022)
   - Derived from Wikipedia sentences

Definition: An mRNA vaccine uses a copy of a molecule called messenger RNA to produce an immune response.
Sentence: mRNA vaccines do not affect or reprogram [MASK].
Entity: mRNA vaccine
Year: 2020
Label: DNA inside the cell

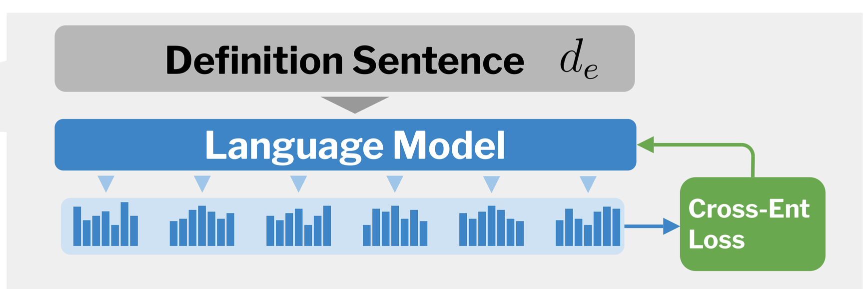
Results

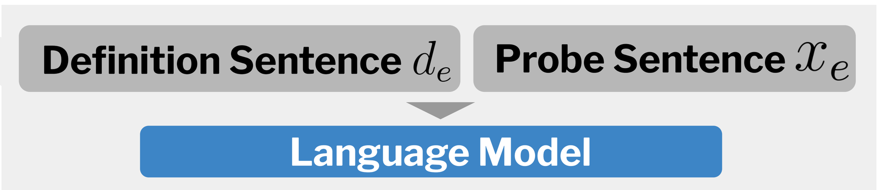




Knowledge Editing Methods

- Standard Finetuning
- MEND (Mitchell et al., 2022)
- ROME (Meng et al., 2022)
- (Baseline) Prepending a definition sentence





Takeaways

- Existing knowledge editing techniques can modify facts but struggle to make inferences based on those facts.
- Prompting baseline (prepending definition) is hard to beat, suggesting that more future research is needed.
- Follow up work that achieves better performance!: **Propagating Knowledge Updates to LMs Through Distillation** (Padmanabhan et al., 2023)

Data is available at github.com/yasumasaonoe/entity_knowledge_propagation