Propagating Knowledge Updates to LMs Through Distillation

TEXAS
The University of Texas at Austin

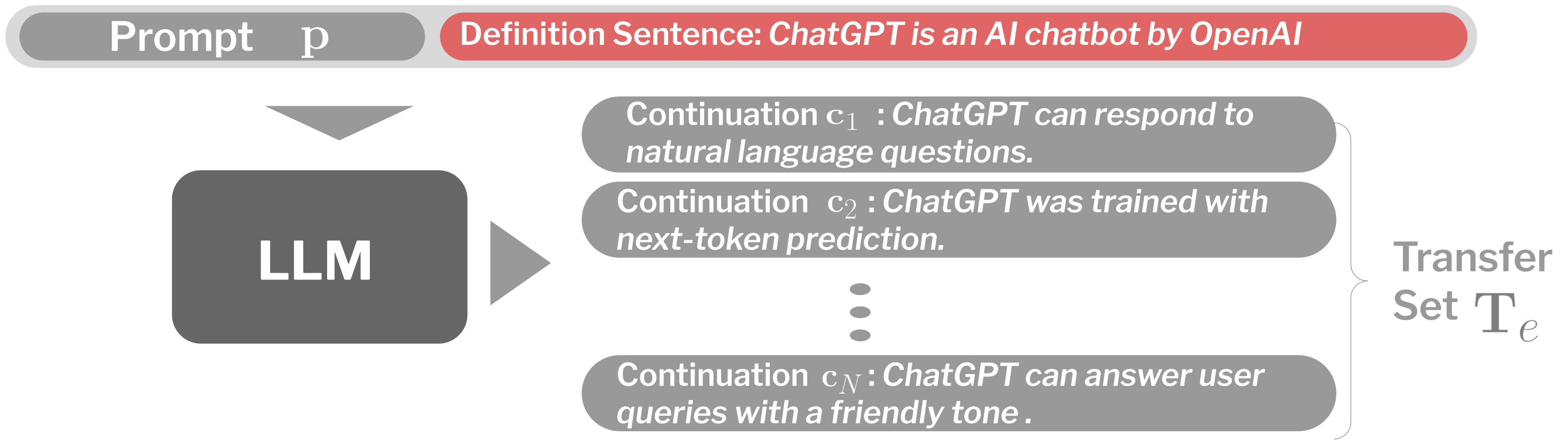
Shankar Padmanabhan, Yasumasa Onoe, Michael J.Q. Zhang, Greg Durrett, Eunsol Choi

NEURAL INFORMATION PROCESSING SYSTEMS

Method

Step 1: Transfer Set Generation

Transfer set: A set of sentences that represent potential inferences given the definition statement. These are continuations of the definition sampled from the language model.



Prompt p | Definition Sentence: *ChatGPT is an AI chatbot by OpenAI*

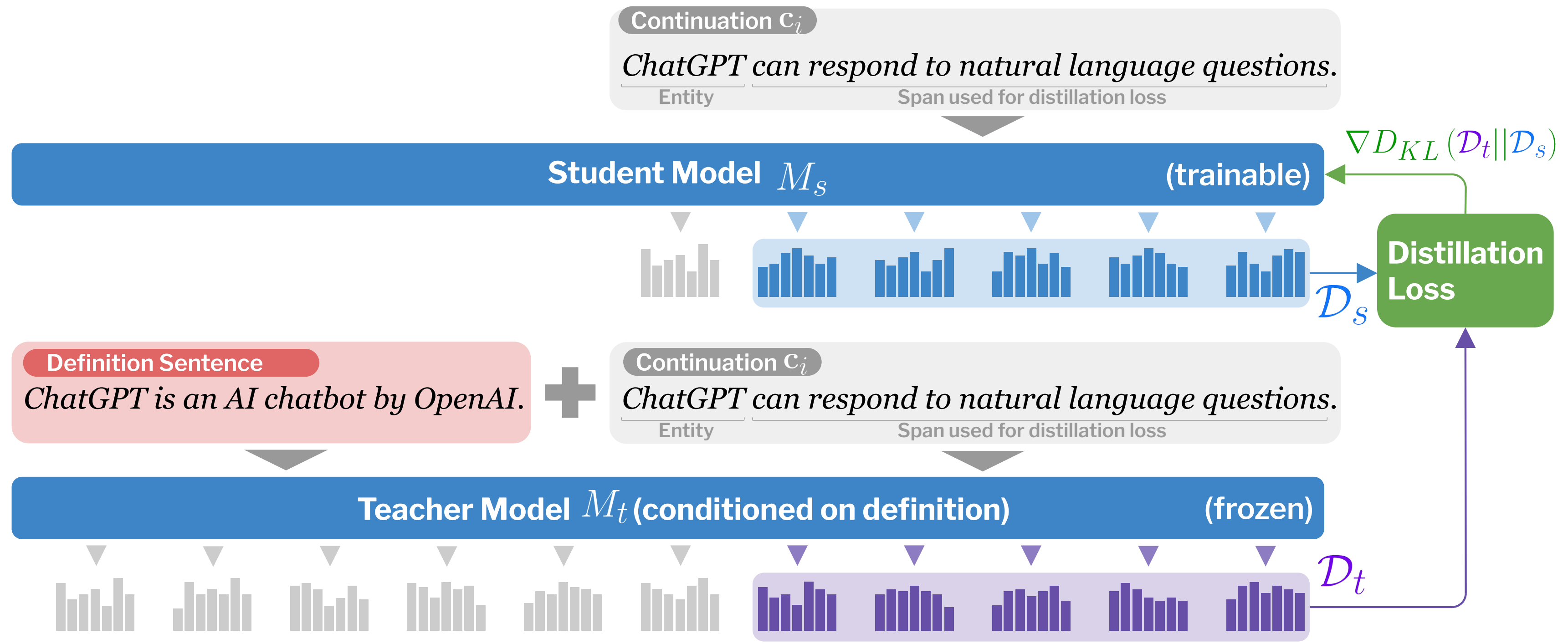
Continuation c_1 : *ChatGPT can respond to natural language questions.*

Continuation c_2 : *ChatGPT was trained with next-token prediction.*

Continuation c_N : *ChatGPT can answer user queries with a friendly tone .*

Step 2: Context Distillation with Transfer Set

The language model (student) is trained to match its distribution when the same model is conditioned on the definition (teacher).



Continuation c_i: *ChatGPT can respond to natural language questions.*

Definition Sentence: *ChatGPT is an AI chatbot by OpenAI.*

Entity
Span used for distillation loss
Student Model M_s
(trainable)
$\nabla D_{KL}(\mathcal{D}_t||\mathcal{D}_s)$
$\mathcal{D}_s$
Distillation Loss
Teacher Model M_t (conditioned on definition)
(frozen)
$\mathcal{D}_t$

Experimental Setup

Datasets:

Entity Inferences (Onoe et al., 2023)

- Manually crafted probe sentences using templates
- Measure **accuracy** on cloze span

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

Entity Cloze By Date (ECBD, Onoe et al., 2022)

- Derived from Wikipedia sentences
- Measure **perplexity** on cloze span

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

Evaluation Metrics

- **Edit Efficacy:** The success of the model edit at *propagating* the injected knowledge
- **Edit Specificity:** The updated models performance on queries about unrelated entities. Ideally, this should not change post-update.

Task: Knowledge Propagation

Prior work edits the parameters of LMs to update their knowledge ("*Rishi Sunak is now the British PM*") and evaluates these updates directly ("*Who is the PM of the UK?*"). We focus on **propagation** of injected knowledge, testing whether LMs can make inferences based on injected facts.

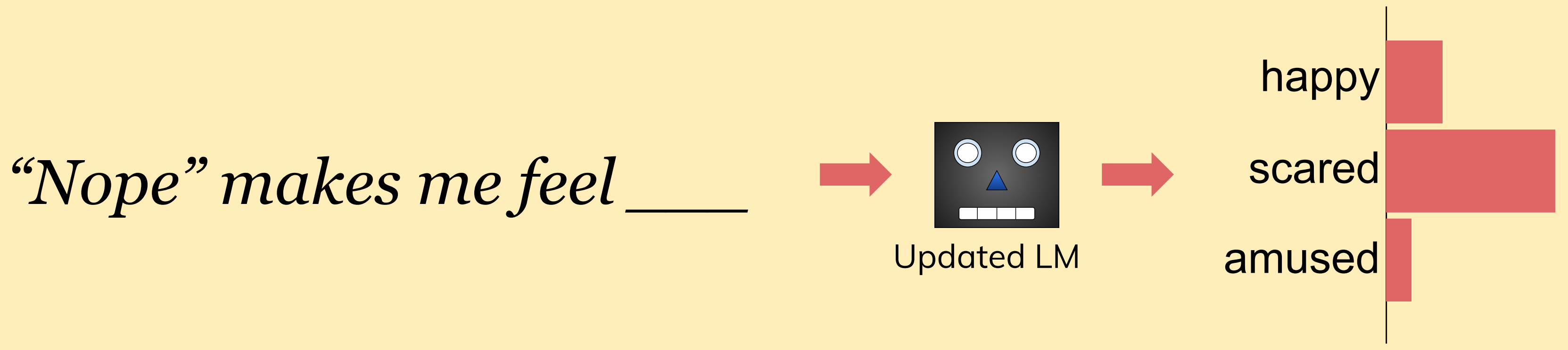
We develop a distillation-based knowledge editing method that can *propagate* injected knowledge!

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

"Nope" is a 2022 American Western science fiction horror film written, directed, and produced by Jordan Peele.

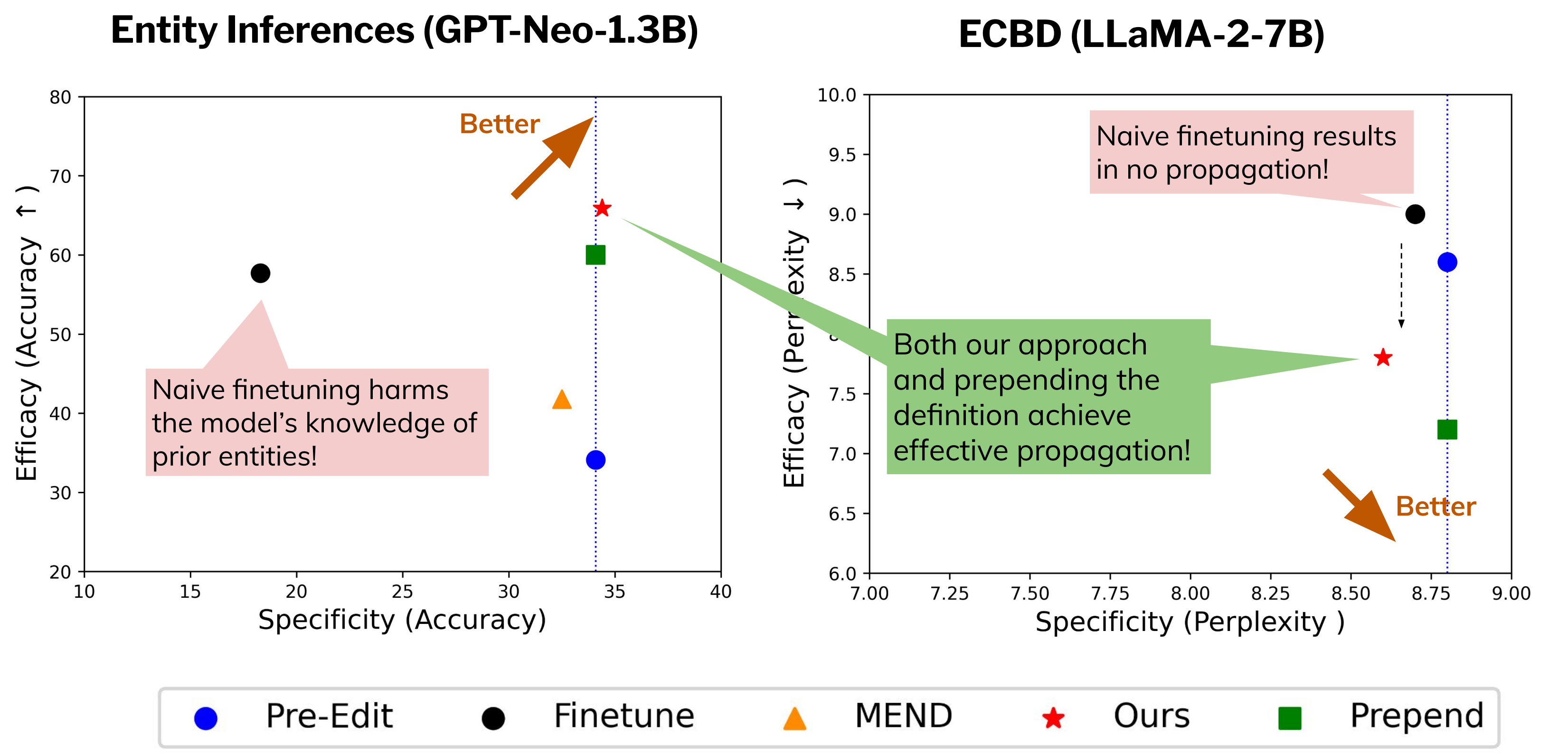
Base LM
Update
Updated LM

The updated LM is evaluated for **propagation** of the knowledge: can the model make inferences based on the fact?



"Nope" makes me feel ____

Results

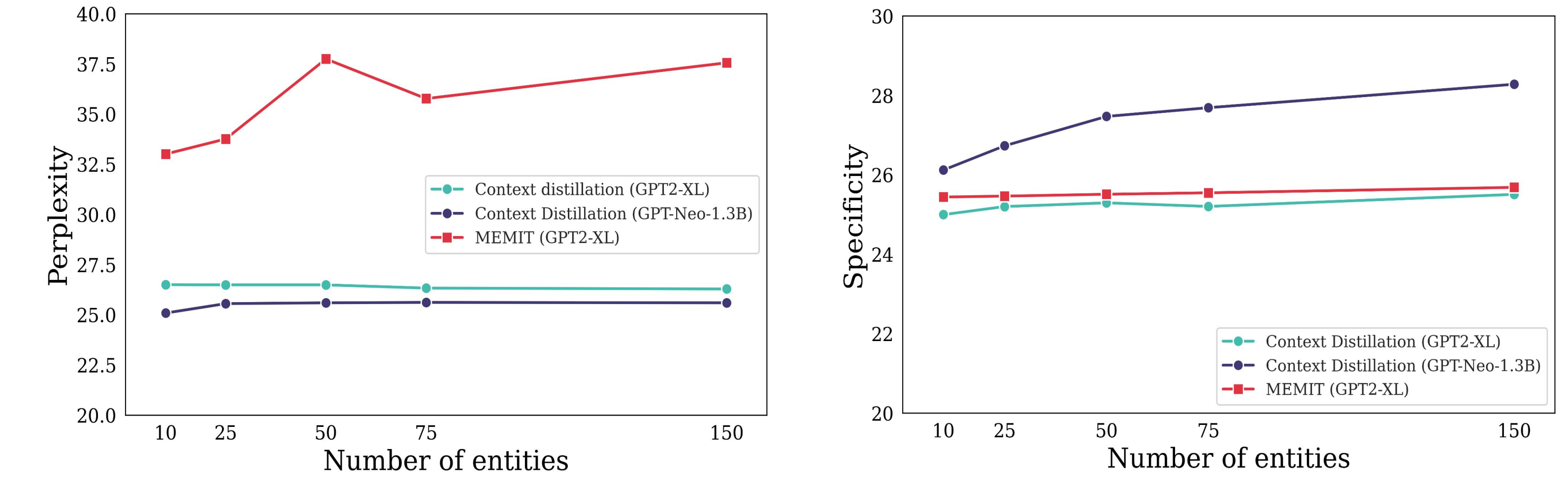


Naive finetuning harms the model's knowledge of prior entities!

Naive finetuning results in no propagation!

Both our approach and prepending the definition achieve effective propagation!

- Our method scales to many entities at once! (ECBD)



- How much does the definition or transfer set matter?

Using the wrong definition hurts performance!

Definition	Transfer Set	Perplexity	Specificity
Random	Correct (GPT-3.5)	33.6 (+2.6)	25.8 (-0.3)
Correct	Random	28.9 (-2.1)	26.6 (+0.5)
Correct	Random + Ent str.	26.7 (-4.3)	25.7 (-0.4)
Correct	Correct (GPT-3.5)	25.3 (-5.7)	25.6 (-0.5)
Correct	Correct (Base)	26.0 (-5.0)	25.9 (-0.2)

Simply appending the correct entity name improves performance!

Using the base model as the generator is nearly as good as using a SOTA model!

- More in the paper:
  - Does the model memorize the definition during distillation?
  - What *kind* of updates is distillation capable of making?
  - How does our method perform on Counterfact? Propagates to related info, but not as good as ROME. (However, ROME performs very poorly on ECBD.)

Code available at github.com/shankarp8/knowledge_distillation