# Experimental Setup

**Method**

**Step 1: Transfer Set Generation**

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

**Step 2: Context Distillation with Transfer Set**

- **Method:**
  - **Edit Specificity:** The updated models performance on queries about unrelated entities.
  - **Edit Efficacy:** The success of the model edit at propagating knowledge updates to LMs through distillation.

**Datasets:**

- **Entity Inferences (GPT-Neo-1.3B)**
- **ECBD (LLaMA-2-7B)**

**Evaluation Metrics**

- **Specificity:** The updated LM makes inferences based on the fact.
- **Accuracy:** The updated LM makes inferences based on the fact.

**Results**

- **Entity Inferences (GPT-Neo-1.3B)**
  - **Specificity:** 0.98
  - **Accuracy:** 0.97

- **ECBD (LLaMA-2-7B)**
  - **Specificity:** 0.95
  - **Accuracy:** 0.96

**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.)

**Results**

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

**Code available at:** github.com/shankarp8/knowledge_distillation

---

**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.

**Example**:

- **Definition:** Hurricane Nana was a minimal Category 1 hurricane that destroyed my house.
- **Transfer Set:** ChatGPT was trained with next-token prediction.
- **Continuation:** Hurricane Nana is a 2022 American Western science fiction horror film written, directed, and produced by Jordan Peele.

**Results**

- **Specificity:** 0.98
- **Accuracy:** 0.97

**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

---

**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.

**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).**

**Experimental Setup**

**Datasets:**

- **Entity Inferences:**
  - **Definition Sentence:** ChatGPT is an AI chatbot by OpenAI.
  - **Label:**
    - **Entity:** Messenger RNA to produce an immune response.
    - **Year:** 2020

**Evaluation Metrics**

- **Perplexity:** Propagating knowledge updates to LMs through distillation.
- **Optimality:** The updated LM performs very poorly on ECBD.

**Results**

- **Entity Inferences:**
  - **Specificity:** 0.98
  - **Accuracy:** 0.97

**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

---

**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.

**Example**:

- **Definition:** Hurricane Nana was a minimal Category 1 hurricane that destroyed my house.
- **Transfer Set:** ChatGPT was trained with next-token prediction.
- **Continuation:** Hurricane Nana is a 2022 American Western science fiction horror film written, directed, and produced by Jordan Peele.

**Results**

- **Specificity:** 0.98
- **Accuracy:** 0.97

**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