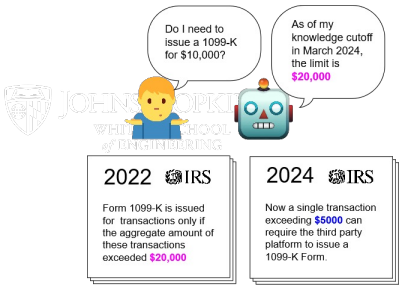
Dated Data: Tracing Knowledge Cutoffs in Large Language Models

Jeffrey Cheng, Marc Marone,
Orion Weller, Dawn Lawrie,
Daniel Khashabi, Benjamin Van Durme

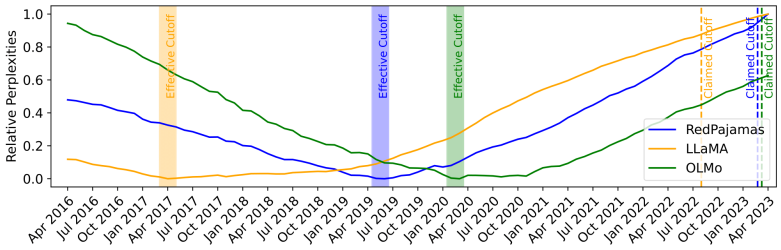
JOHNS HOPKINS

Introduction

- LLM creators often do not elect to release training data, instead providing a *reported cutoff* date.
- Is this knowledge cutoff the same for each of its included resources (e.g. Wikipedia, ArXiv, Github)?
- Does this knowledge cutoff match the model's knowledge of the resource, or *effective cutoff*?



Claimed and Effective Cutoffs

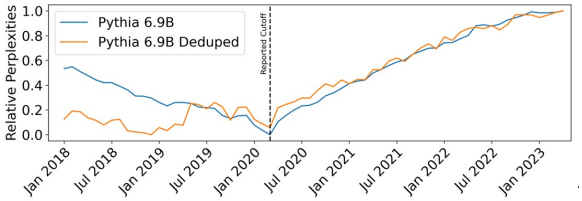


- We propose a simple method to determine *effective cutoffs* without needing access to pre-training data, creating long spanning (2016-2023) datasets.
- Our datasets consist of versions of Wikipedia and NYT documents, and we take the time at which perplexities are minimized to be the *effective cutoff* of that resource.
- We measure *effective cutoffs* across a variety of language models and show that there exists drastic mismatches with *reported cutoffs*.

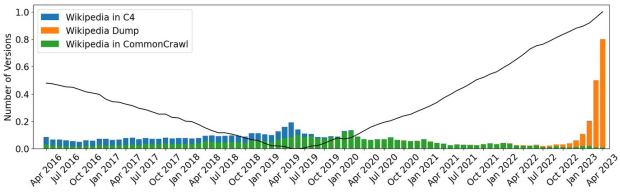
Can you trust reported knowledge cutoffs?

Duplication

- Despite it being common practice to deduplicate pre-training datasets, we empirically find many duplicates in most pre-training datasets.
- We consider the Pythia suite, whose *effective cutoffs* match the *reported cutoffs* in part due to the purposeful upsampling of document versions at the *reported cutoff* date.
- We confirm duplicate documents affect *effective cutoffs* by considering Pythia-deduped, which removes the upsampled documents.



CommonCrawl Misalignments



- Most modern LLMs are trained on CommonCrawl data, and our analysis reveals that a non-trivial amount of data inside each dump is old data.
- We concretely show this for RedPajamas; the majority of its CommonCrawl dumps occur after 2020 yet most of its Wikipedia versions are from before 2020.

We identify two reasons that contribute to the temporal mismatch of a language model's reported and effective cutoff:

(1) failures of deduplication pipelines to control for semantic duplicates and

(2) the use of newer CommonCrawl dumps to provide updated information when they include significant amounts of older data