Mitigating Temporal Misalignment by Discarding Outdated Facts

Michael J.Q. Zhang and Eunsol Choi

The University of Texas at Austin

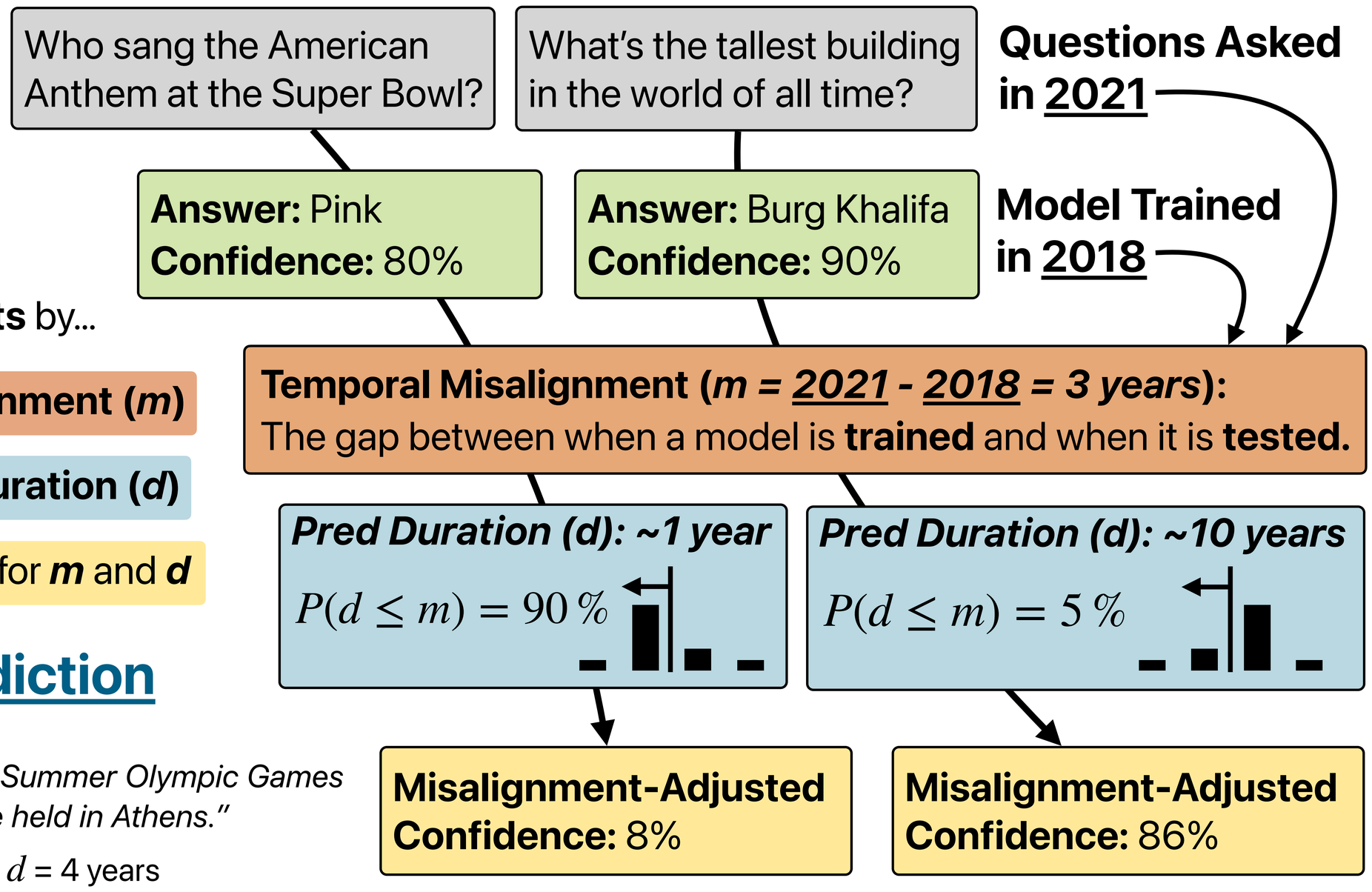
{mjqzhang,eunsol}@utexas.edu

LMs are prone to...

reciting **outdated facts** with **high confidence**

We **discard outdated facts** by...

1. Measuring the **misalignment (*m*)**
2. Predicting the **fact's duration (*d*)**
3. Adjusting **confidence** for ***m*** and ***d***



Fact Duration Prediction

Given a fact. f = *"The last Summer Olympic Games were held in Athens."*

Predict its duration d = 4 years

Fact Duration Modeling:

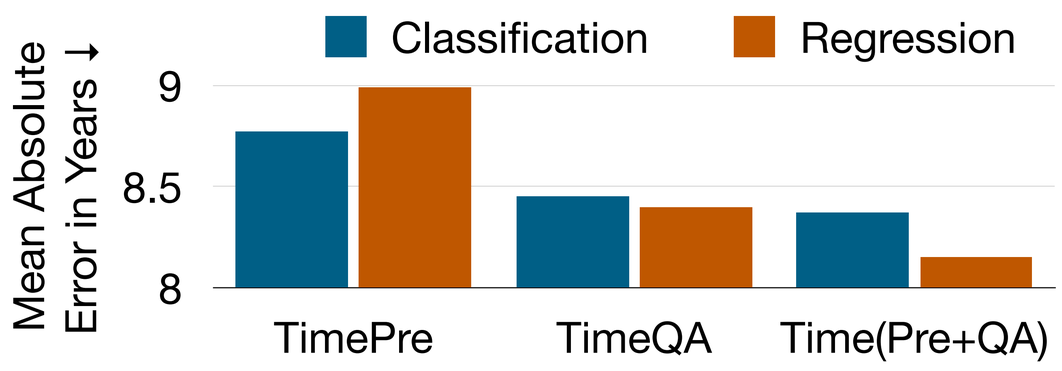
We train **regression** and **classification** BERT models

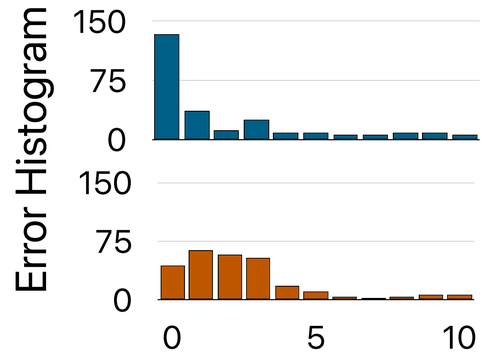
Training w/ Distant Supervision:

TimePre: Duration-related **news text**

TimeQA: Temporal **knowledge-bases**

Fact Duration Evaluation: Facts from SituatedQA





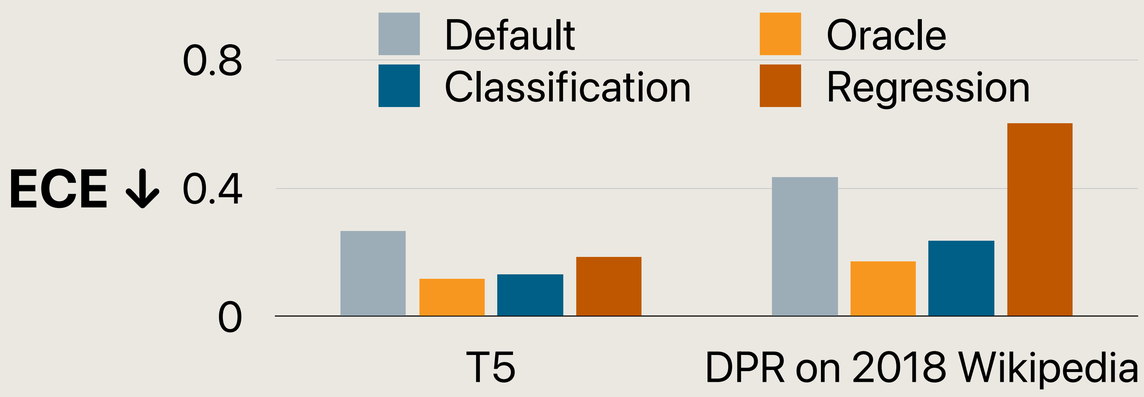
Can we STOP models from predicting outdated facts?

Adjusting Confidence w/ Predicted Durations:

We use **T5** and **DPR**, both pretrained and finetuned in **2018**

We adjust confidence for each fact duration model using:

- **Regression:** Zero confidence if $m > d$
- **Classification:** Use CDF to scale confidence by $P(m > d)$



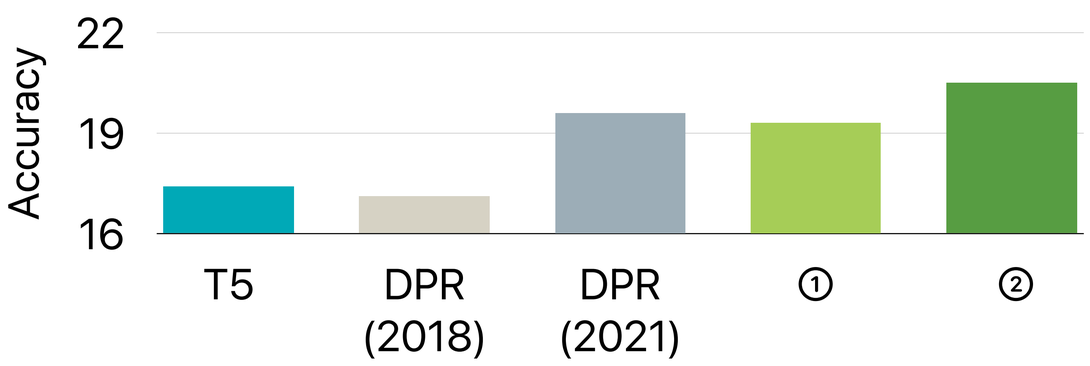
Beyond Calibration: Adaptive Inference

Hybrid Systems — ① T5 + DPR (2021)

- Use a cheaper closed book system when possible

Heterogeneous Retrieval —② DPR (2018 + 2021)

- Simulate retrieval over articles authored on different dates



"As of my last knowledge update..." — GPT-4

We find that GPT-4 **over-abstains**

- GPT-4 abstains on **95% of facts that have changed** between 2018 and 2021 and **79% of facts that have not changed.**

Does GPT-4 abstain on questions with durations of ≤ 2 years?

- GPT-4's knowledge cutoff was in 2021. We evaluate in 2023.