

Mitigating Temporal Misalignment by Discarding Outdated Facts



TEXAS
The University of Texas at Austin

Michael J.Q. Zhang and Eunsol Choi

{mjqzhang, eunsol}@utexas.edu

LMs are prone to...

reciting **outdated facts**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

Who sang the American Anthem at the Super Bowl?

Answer: Pink
Confidence: 80%

What's the tallest building in the world of all time?

Answer: Burg Khalifa
Confidence: 90%

Questions Asked in **2021**

Model Trained in **2018**

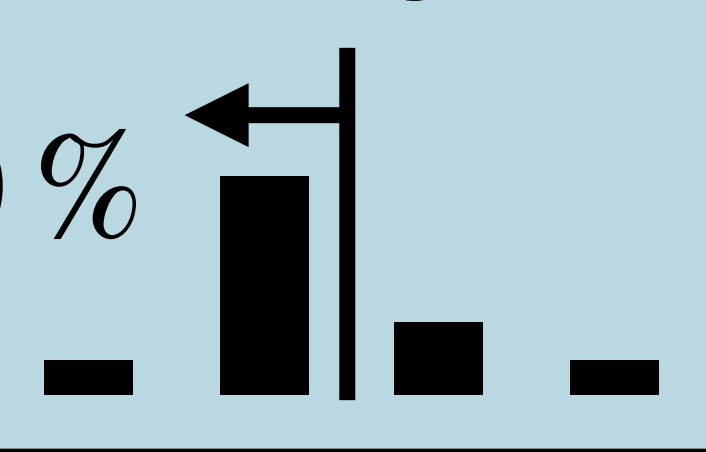
We discard outdated facts by...

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

Temporal Misalignment ($m = 2021 - 2018 = 3$ years):
The gap between when a model is **trained** and when it is **tested**.

Pred Duration (d): ~1 year

$$P(d \leq m) = 90\%$$



Pred Duration (d): ~10 years

$$P(d \leq m) = 5\%$$



Fact Duration Prediction

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

Predict its duration

$d = 4$ years

Misalignment-Adjusted Confidence: 8%

Misalignment-Adjusted Confidence: 86%

Fact Duration Modeling:

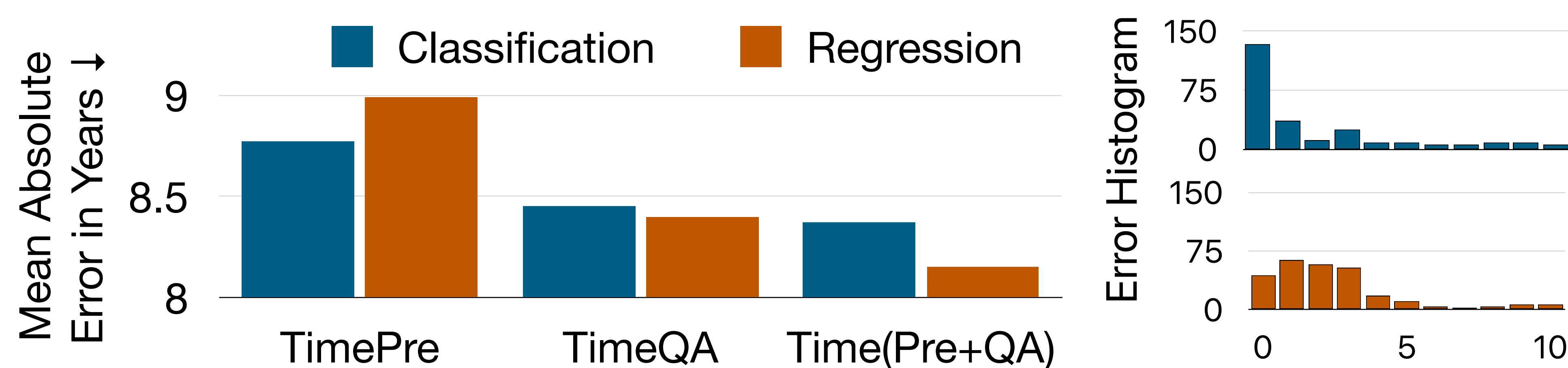
We train **regression** and **classification** BERT models

Training w/ Distant Supervision:

TimePre: Duration-related news text

TimeQA: Temporal knowledge-bases

Fact Duration Evaluation: Facts from SituatedQA



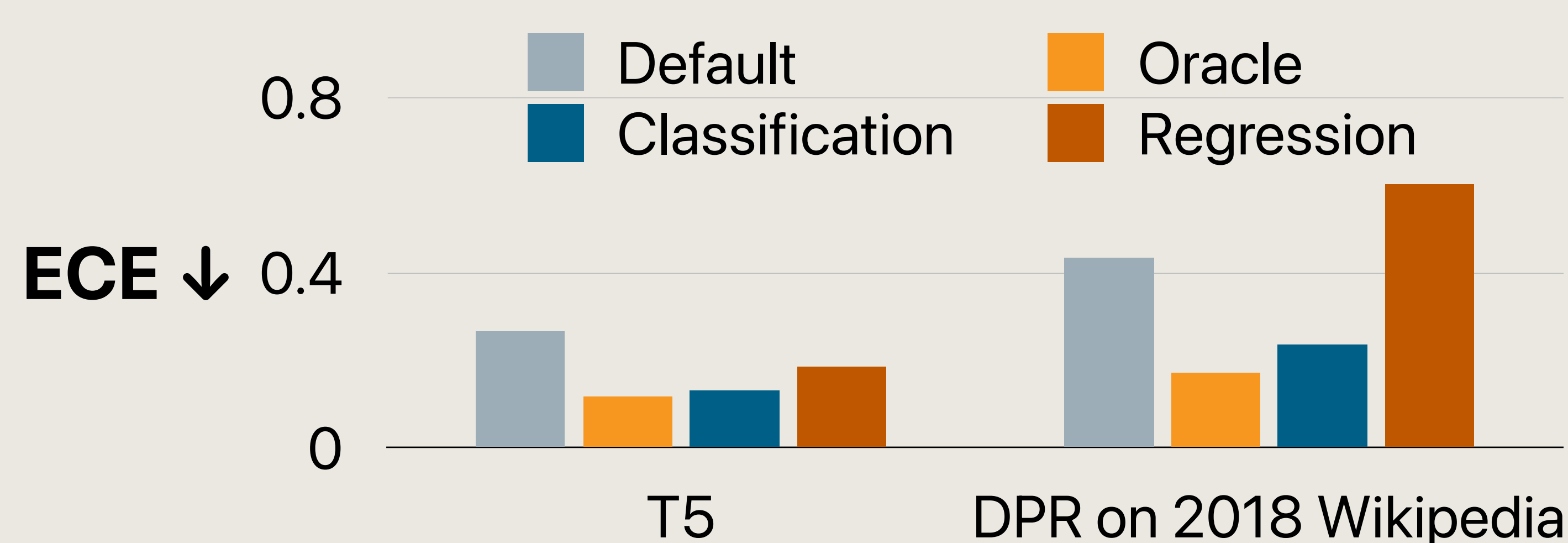
Can we **STOP** models from predicting outdated facts?

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 $P(m > d)$



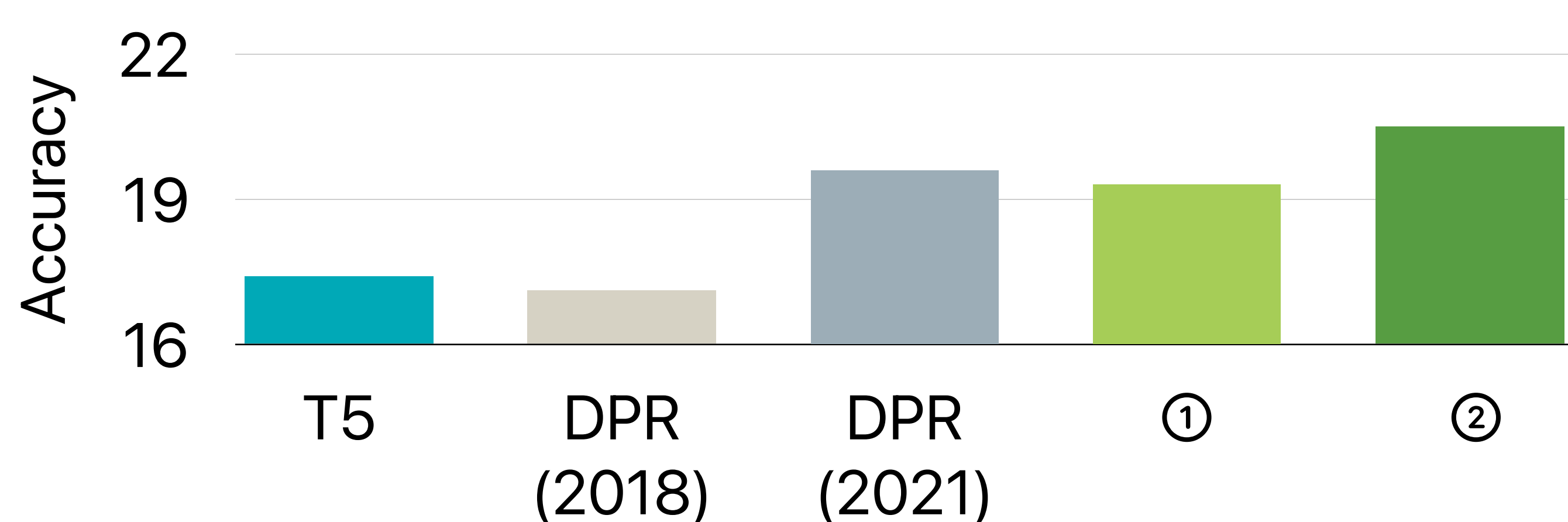
Beyond Calibration: Adaptive Inference

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.

