DeepMind

StreamingQA

A Benchmark for Adaptation to New Knowledge over Time in Question Answering Models

Adam Liška*, **Tomáš Kočiský***♠, Elena Gribovskaya*♠, Tayfun Terzi*,
Eren Sezener, Devang Agrawal, Cyprien Masson d'Autume, Tim Scholtes, Manzil Zaheer, Susannah Young,
Ellen Gilsenan-McMahon, Sophia Austin, Phil Blunsom, Angeliki Lazaridou

* Equal contribution in random order ♠Project Leads

How to measure adaptation to and forgetting of new knowledge?

QA is useful for interrogating models about their language understanding, knowledge, reasoning, and for various knowledge-oriented applications (personal assistants or web search).

The questions people would ask in practice can be about:

Any point in the history (e.g., 4 years ago)

<u>Question Date:</u> Sunday, Apr 12, 2020
<u>Question:</u>
In November 2016, which Netflix series set in the United Kingdom was said to be “the most expensive television series ever”?

Recent events (last few weeks or days)

<u>Question Date:</u> Monday, Feb 24, 2020
<u>Question:</u>
How many countries have committed to the net zero target as of today's date?

As the world and knowledge evolve, we need our QA models to adapt to such new information, to not forget the past, and to maintain an up-to-date world model.

How to measure adaptation to and forgetting of new knowledge?

To be able to ask such questions **we need a dataset with temporal grounding of both the**

- **Questions** – with dates when questions were asked,
- **Knowledge** – with when the articles were published.

No current dataset allows us to do this!

StreamingQA Dataset

github.com/deepmind/streamingqa

Knowledge Corpus:

14 years (2007–2020) of English WMT news with publication dates. (11M articles / 48M passages for retrieval)

Example from the dataset:

Question Date: Sunday, April 12, 2020
Question: In November 2016, which Netflix series set in the United Kingdom was said to be "the most expensive television series ever"?

Plus:

- 3 reference answers
- Gold evidence article + publication date

StreamingQA Dataset

github.com/deepmind/streamingqa

Knowledge Corpus:

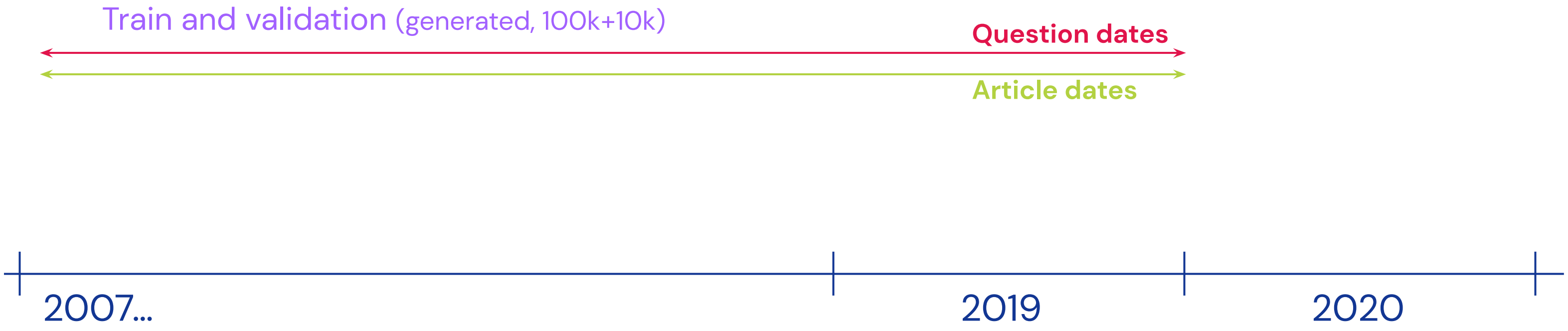
14 years (2007–2020) of English WMT news
with publication dates. (11M articles / 48M passages for retrieval)

Example from the dataset:

Question Date: Sunday, April 12, 2020
Question: In November 2016, which Netflix series set in the United Kingdom was said to be "the most expensive television series ever"?

Plus:

- 3 reference answers
- Gold evidence article + publication date



StreamingQA Dataset

github.com/deepmind/streamingqa

Knowledge Corpus:

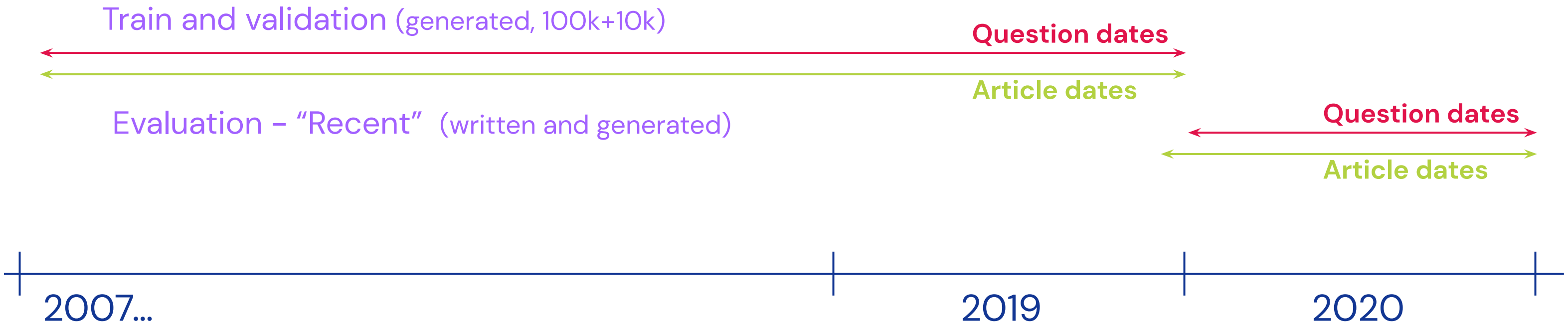
14 years (2007–2020) of English WMT news
with publication dates. (11M articles / 48M passages for retrieval)

Example from the dataset:

Question Date: Sunday, April 12, 2020
Question: In November 2016, which Netflix series set in the United Kingdom was said to be "the most expensive television series ever"?

Plus:

- 3 reference answers
- Gold evidence article
+ publication date



StreamingQA Dataset

github.com/deepmind/streamingqa

Knowledge Corpus:

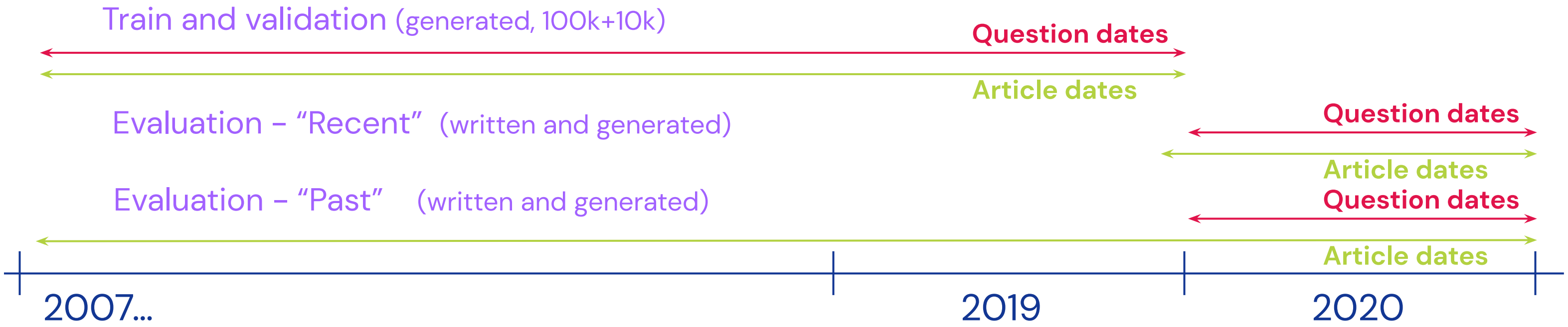
14 years (2007–2020) of English WMT news
with publication dates. (11M articles / 48M passages for retrieval)

Example from the dataset:

<u>Question Date:</u> Sunday, April 12, 2020
<u>Question:</u> In November 2016, which Netflix series set in the United Kingdom was said to be "the most expensive television series ever"?

Plus:

- 3 reference answers
- Gold evidence article
  + publication date



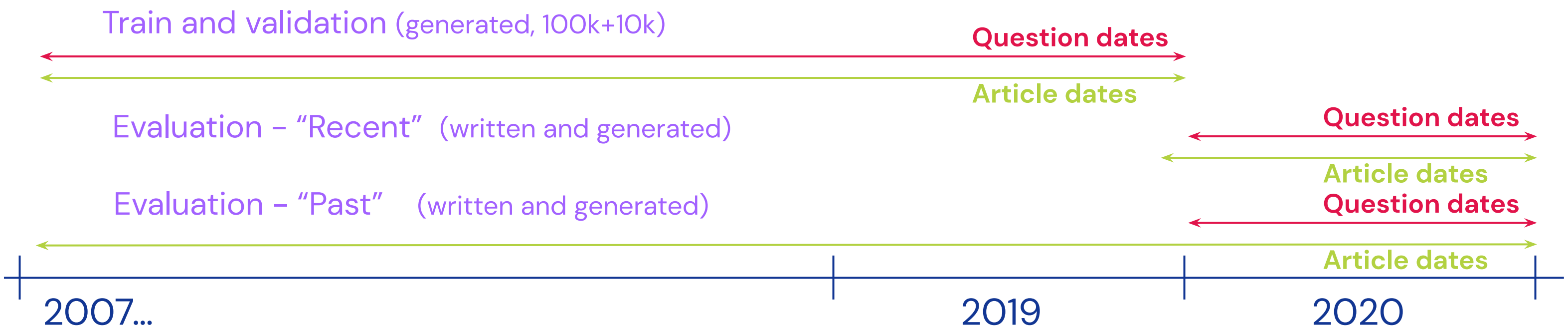
StreamingQA Dataset

github.com/deepmind/streamingqa

We **generated questions** through few-shot prompting of a large LM (which was not used for our experiments here).

Written questions are written by human annotators (given a news article and a desired question date).

All evaluation data (generated and written) is **human filtered**.



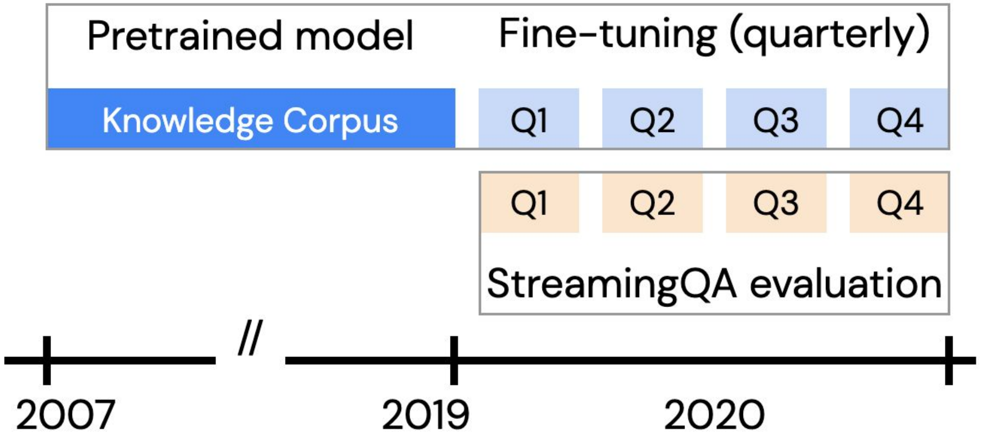
A Streaming Task

This temporal grounding allows us to control the knowledge in the model
and evaluate adaptation and forgetting over time, as the world evolves.

Every quarter:

Add new knowledge

And evaluate on new questions about recent and past events.



The Underlying LMs

Three underlying LMs:

"**Retrained**" model

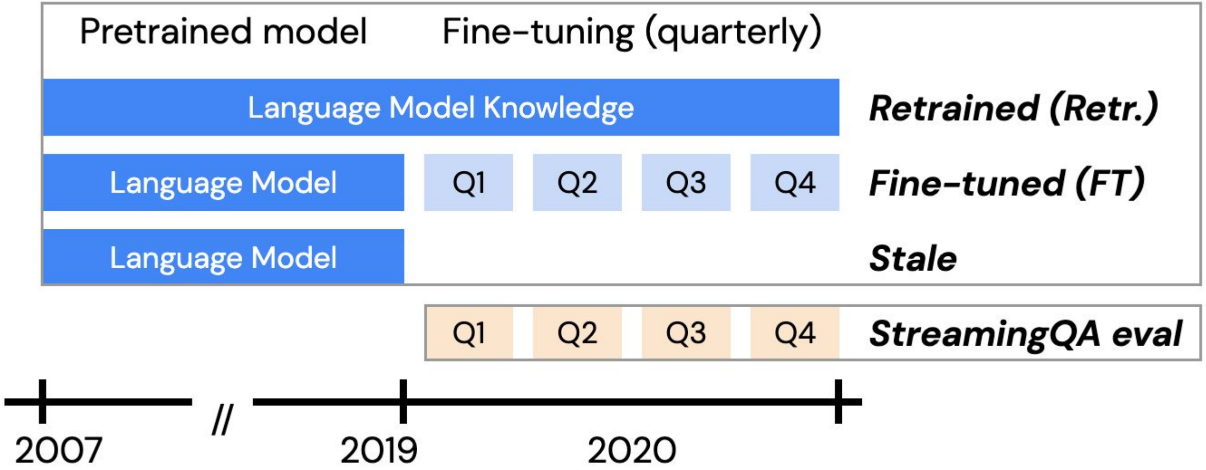
with knowledge of the entire period from 2007 to including 2020.

"**Fine-tuned**" models

at every quarter of 2020.

"**Stale**" model

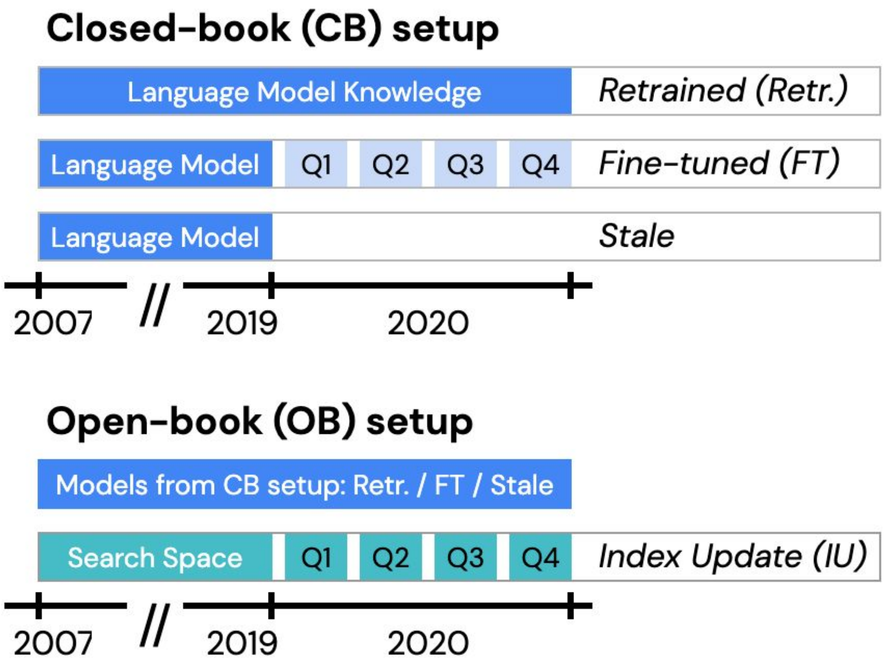
without the 2020 recent evaluation knowledge.



Closed-book and Open-book Models

Closed-book QA models are based on these LMs and fine-tuned for QA.

Open-book retrieval QA models use the same LMs and additionally add new articles to the search space quarterly (IU).



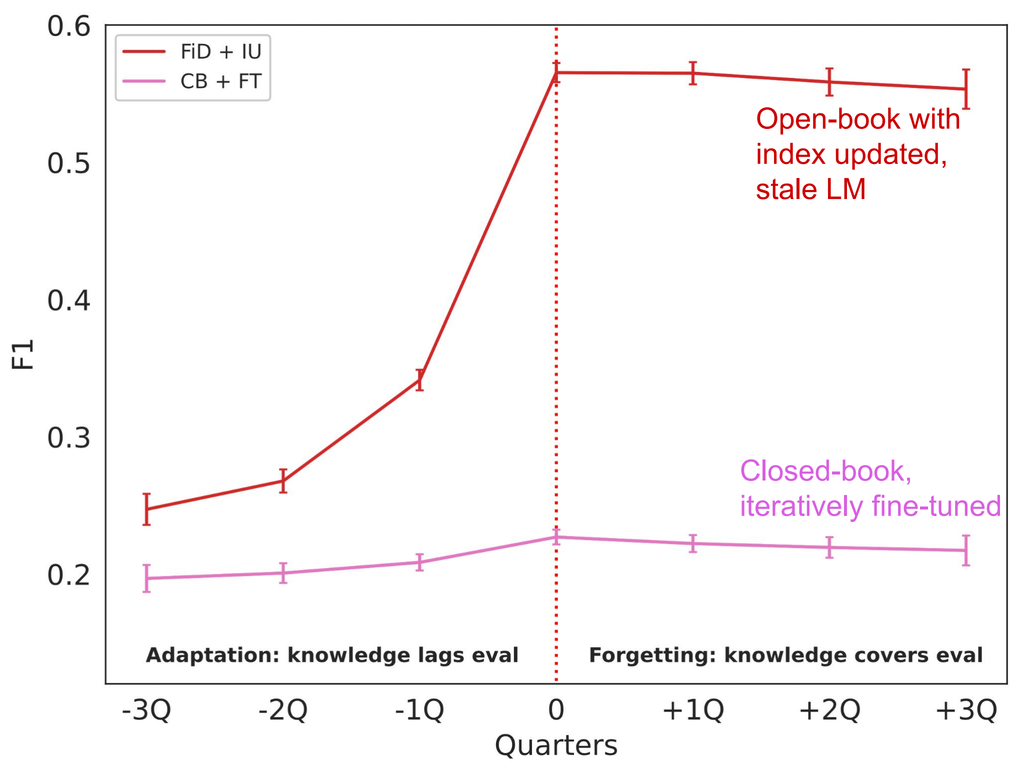
Temporal Lag Evaluation

To measure adaptation to new information and forgetting, we investigate the model performance for varying **temporal lag** between:

- knowledge in the underlying LM and
- the question date.

–1Q — model is adapting to new knowledge, knowledge lags 1Q behind when the question was asked.

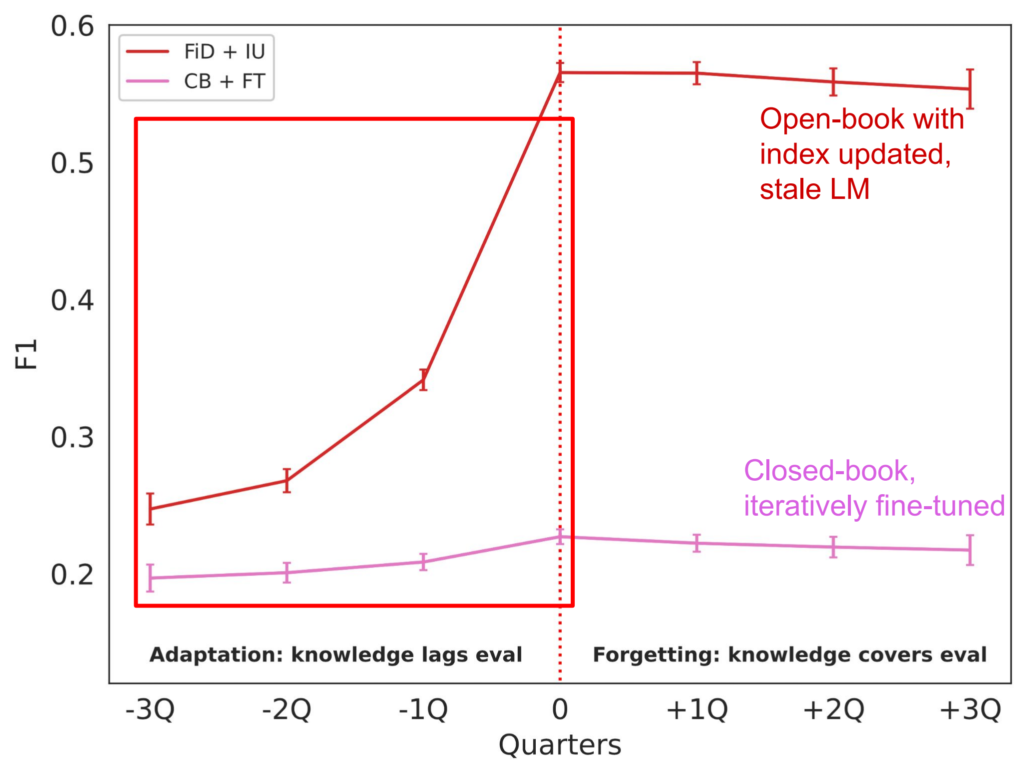
+2Q — model has additional knowledge and may be forgetting the past.



Adaptation vs Forgetting: Generated questions about Recent events.

As the model acquires the necessary knowledge, from −3Q to 0Q:

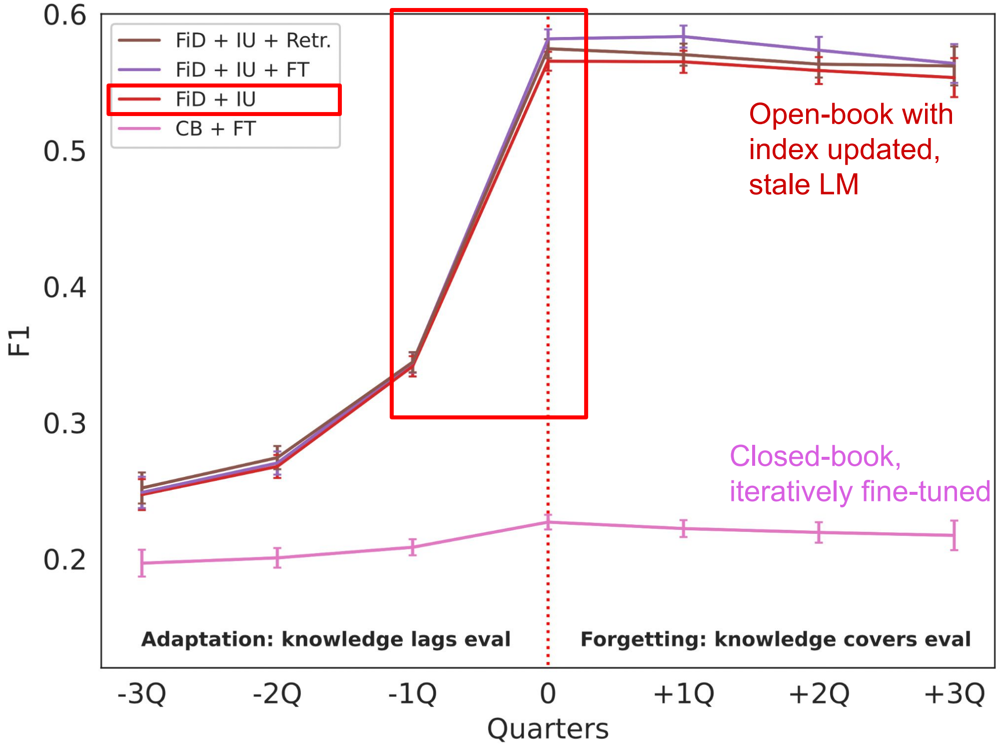
The model is performing much better, so **our dataset indeed requires this recent new knowledge at 0Q**.



Adaptation vs Forgetting: Generated questions about Recent events.

-1Q to 0Q: We see a steep adaptation rate for the open-book (FiD) models.

Just adding articles into the search space performs quite well.

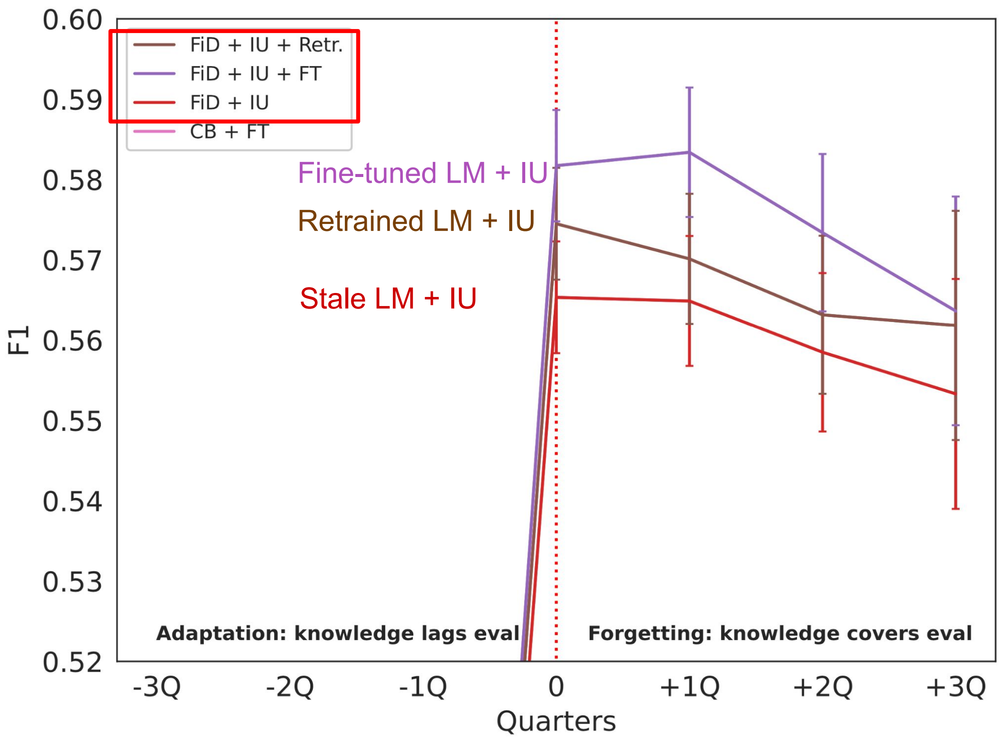


Adaptation vs Forgetting: Generated questions about Recent events.

Next we wanted to know:
Can we update the search index over time and rely on a Stale LM?

Fine-tuning or retraining still further improves the performance.

In our paper, you can find an analysis for which questions does the fine-tuning help.

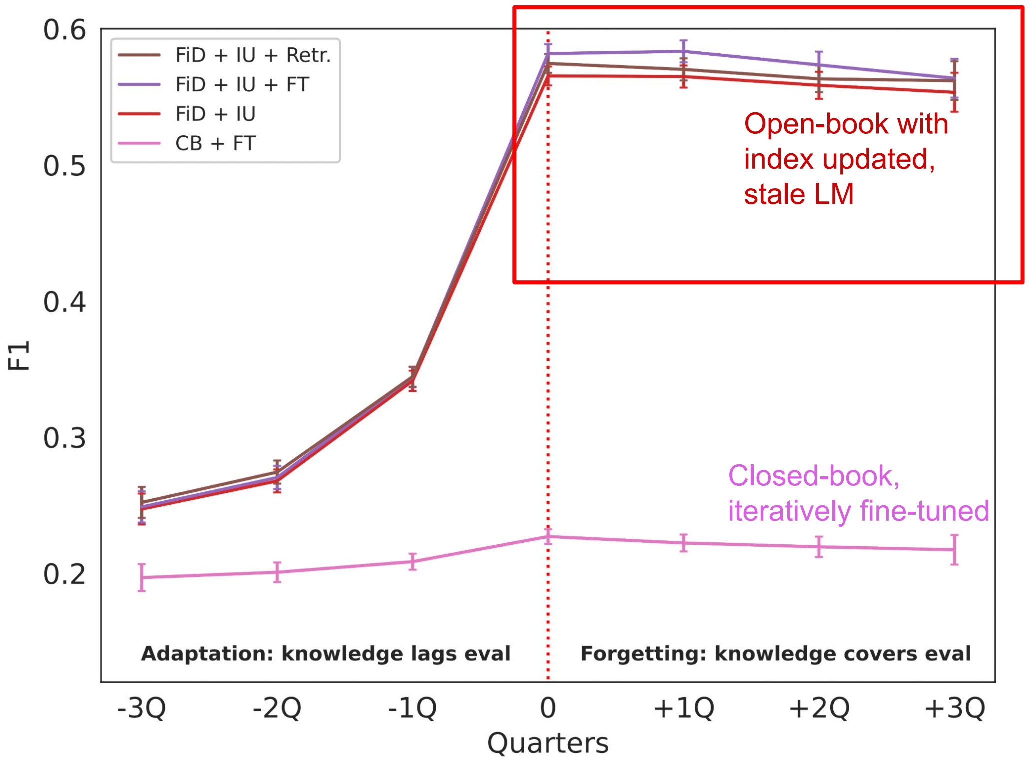


IU = search index updated

Adaptation vs Forgetting: Generated questions about Recent events.

StreamingQA

As we keep fine-tuning, for all model, we see almost no forgetting.



See our paper for

github.com/deepmind/streamingqa

Details of experiments on the other evaluation subsets

Past generated, and recent and past written

Metrics for

The usual QA setup (with temporal context, over news)

One-step continual learning QA setup

More analysis, including toxicity analysis and details of our filtering

Conclusions

github.com/deepmind/streamingqa

To enable a more realistic evaluation of QA models, we introduced the **StreamingQA** dataset with questions **about new knowledge and about all the history**.

We are able to create challenging human written questions, and more scalably, generated questions for this task.

We found that adding new articles into the search space of **open-book models** allows for **quick adaptation**, but fine-tuning and (a lot more costly) retraining (of underlying LMs) further improves performance.

github.com/deepmind/streamingqa

Thank you!