servicenow

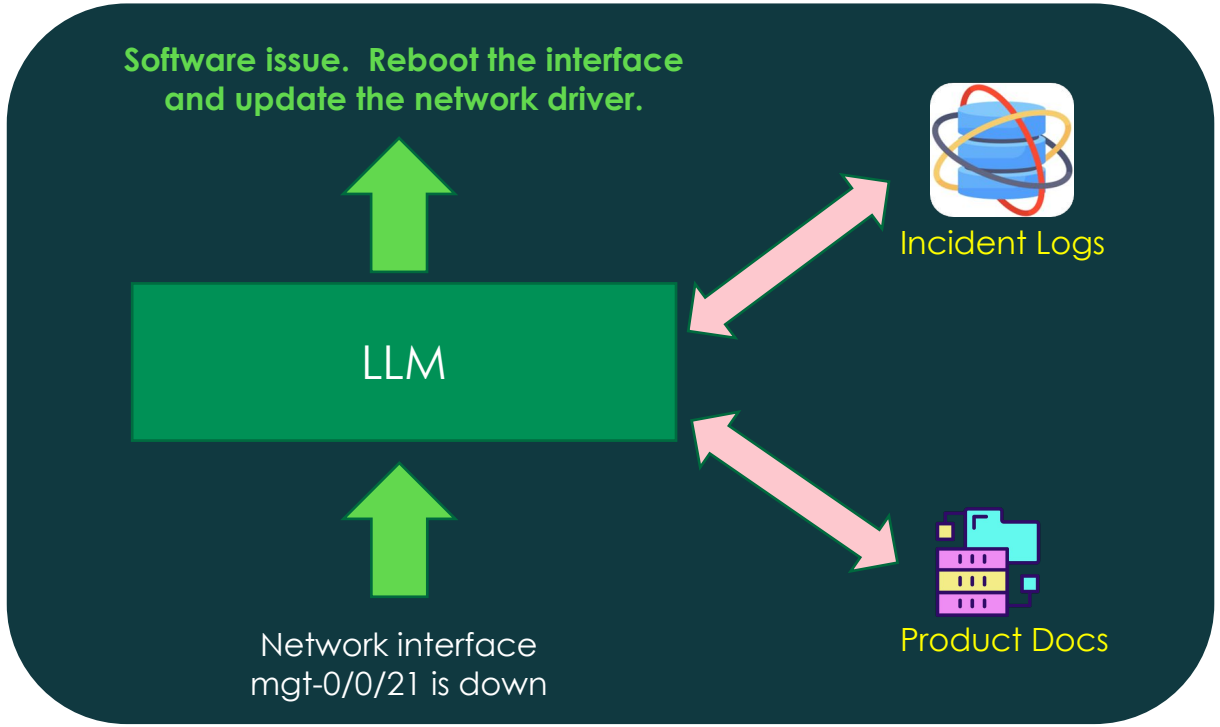
RepLiQA: A Question-Answering Dataset for Benchmarking LLMs on Unseen Reference Content

NeurIPS, 2024

João Monteiro*, Pierre-André Noël, Étienne Marcotte, Sai Rajeswar, Valentina Zantedeschi, David Vázquez, Nicolas Chapados, Christopher Pal, Perouz Taslakian
ServiceNow Research
*Currently at the Autodesk AI Lab

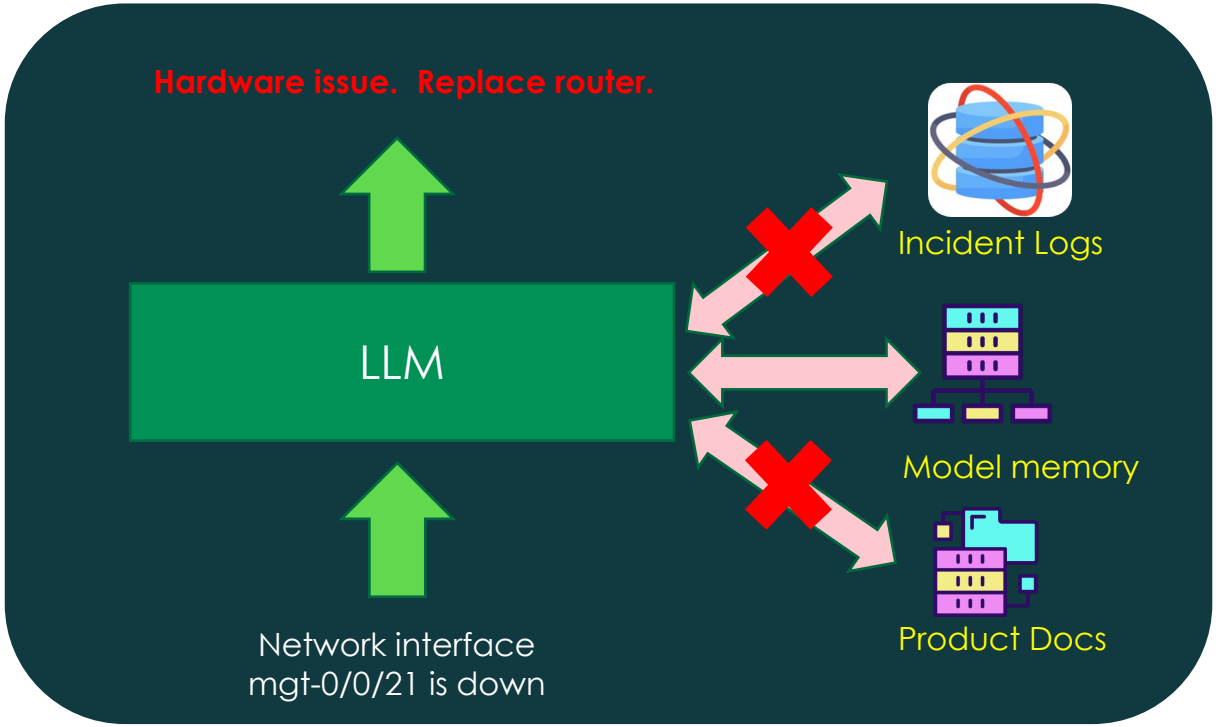
Answering questions based on user-provided contexts

- Question answering helps **resolving issues faster** and enables self-service
- For more accurate outputs, we typically ground on **user-provided information**



Models must be able to access and use user-provided context

- Text generation can be cofounded by **model memory**, obtained during training
- An LLM may ignore the context and **answer based solely on its existing knowledge**



“

How can we test for the ability of different language models to search and effectively use context information provided by the user?

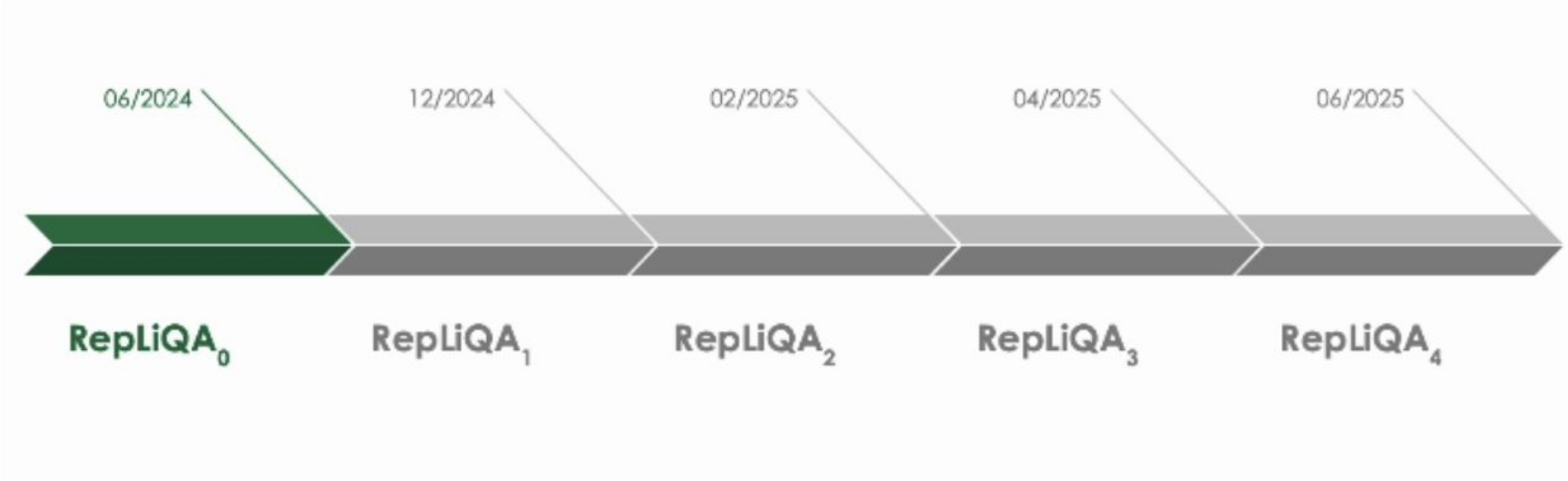
Confounded assessments with existing QA datasets

- Popular benchmarks built using information **openly available** on the web
  - Models retrieving answer from their training memory can obtain good results on these benchmarks…
- For my organization with its **specific knowledge base, policies, logs, etc.,** which model should I use?

We need new benchmarks to distinguish between memorization and the ability to use relevant user-provided context

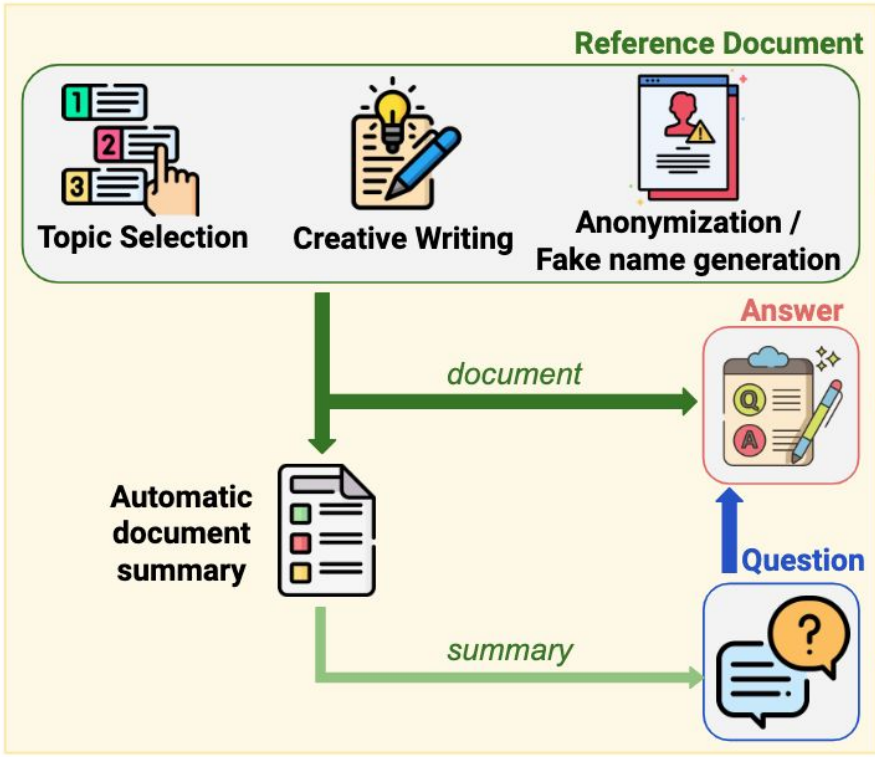
Introducing RepLiQA

- A collection of reference documents, each with 5 question-answer pairs:
  - All feature **fictional scenarios** created by **human content writers**
  - Guarantees that models cannot rely on training memory
- **Scheduled releases of fresh splits** for sustainable rigorous evaluation



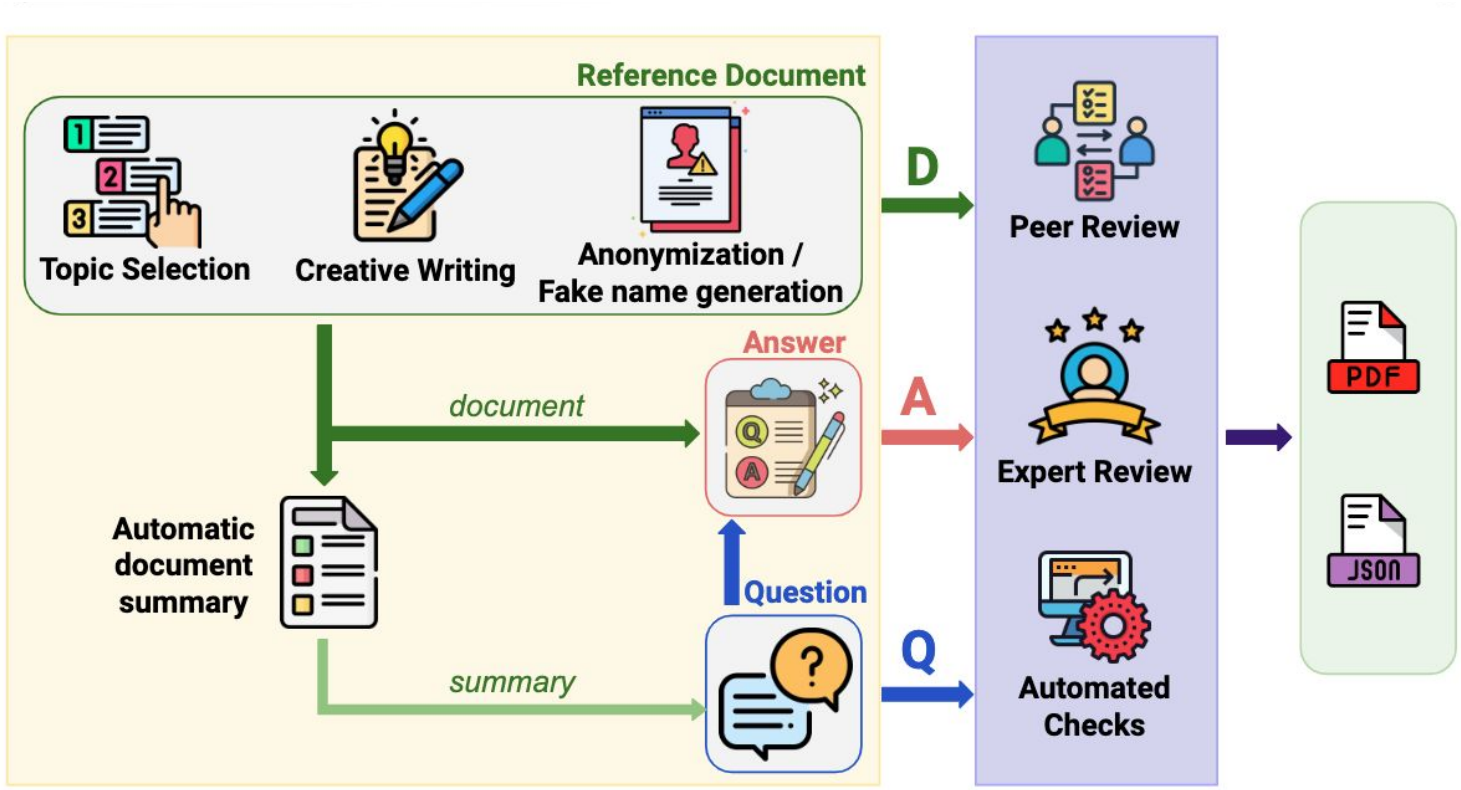
Building RepLiQA

Step 1: Document, question, answer triplets are created by a team of annotators.



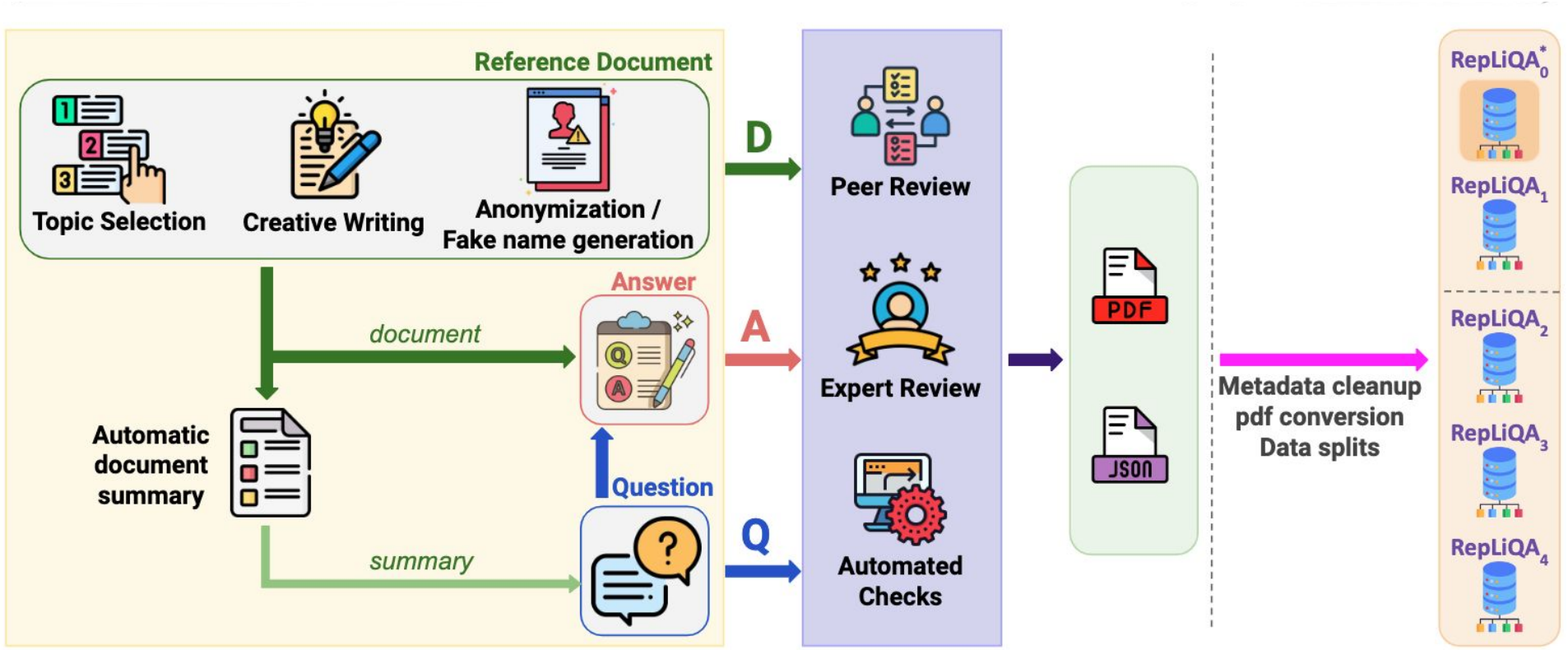
Building RepLiQA

Step 2: Quality controls are carried out to filter out some of the annotations.



Building RepLiQA

Step 3: We perform additional quality controls and filter out some more data. We also split the data for sequential releases.



RepLiQA' fictitious scenarios

Topic: Cybersecurity News

Cybersecurity in Education: The Critical Need for Educator and Staff Vigilance

In an age where technological integration into every facet of life is commonplace, the education sector finds itself grappling with a relatively new but rapidly growing challenge: cyber threats. [...]

The Growing Threat Landscape

[...] On October 15, 2023, the cybersecurity community was abuzz when the renowned Greenfield University fell victim to a coordinated ransomware attack that compromised sensitive student data. [...]

Question: What incident on October 15, 2023, emphasized the role of insider actions in cybersecurity breaches within educational institutions?
Answer: The ransomware attack on Greenfield University.

A realistic incident that never happened.

Real entities may be mentioned, but scenarios do not contradict openly available information.


Greenfield University

Article Talk

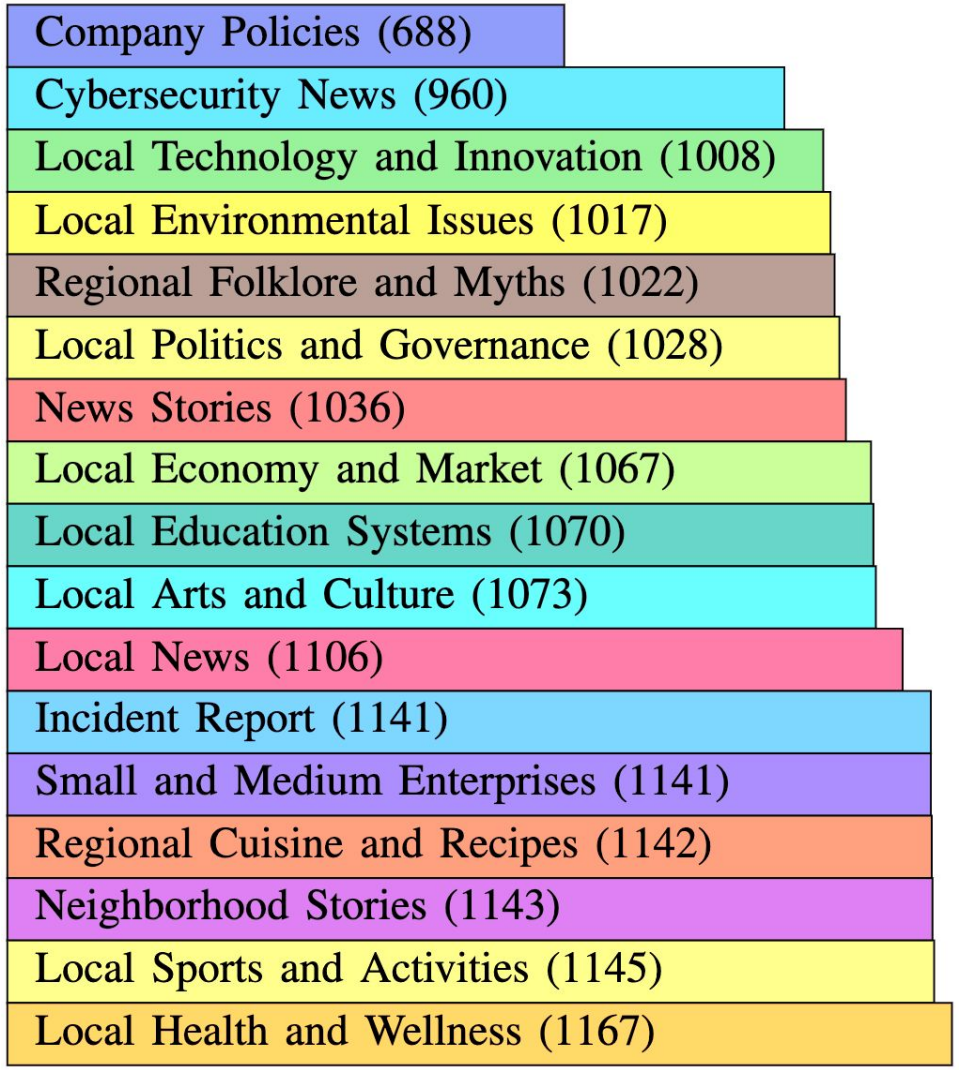
From Wikipedia, the free encyclopedia

Greenfield University is a private university in Kaduna, Kaduna State, Nigeria. It was established in January 2019, and commenced the 2018/19 session in May 2019.

The reference document offers enough information to answer questions.

An overview of RepLiQA

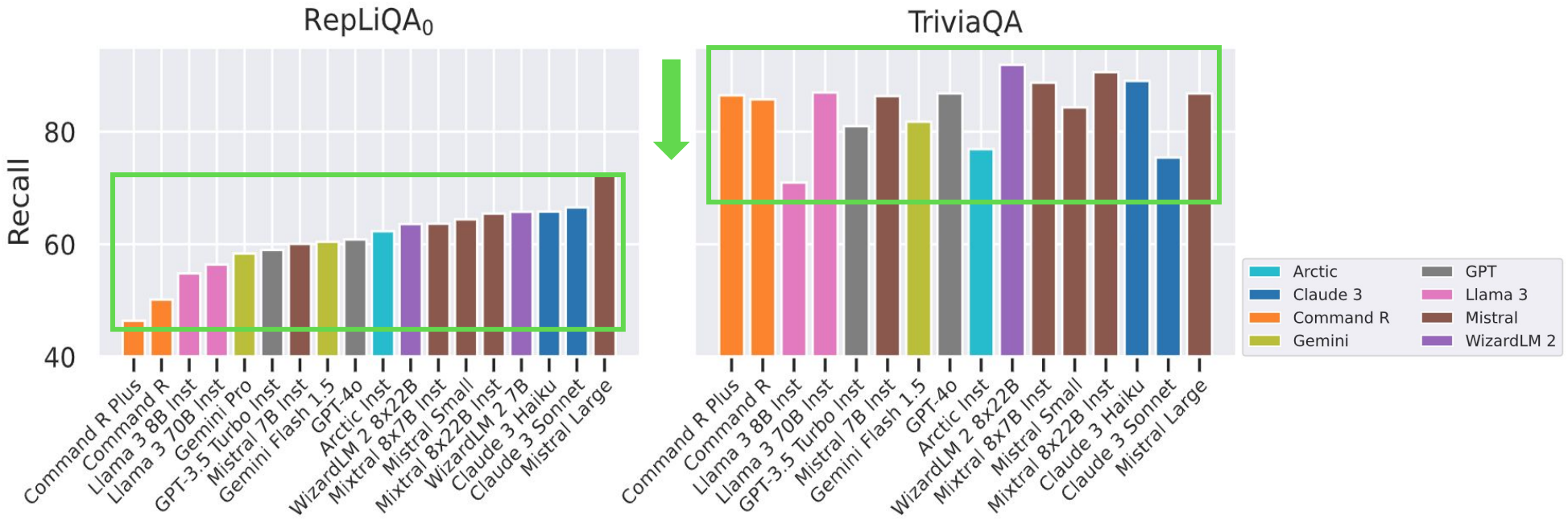
- **17 diverse topics** (somewhat) uniformly covered
- Each document has 5 question-answer pairs
- Some of the questions **cannot be answered based on the reference**
  - We can evaluate if models correctly refuses to reply.



Benchmarking LLMs using RepLiQA

- All models show a **huge degradation in performance** when they cannot rely on memory

  TriviaQA covers **well-known facts**

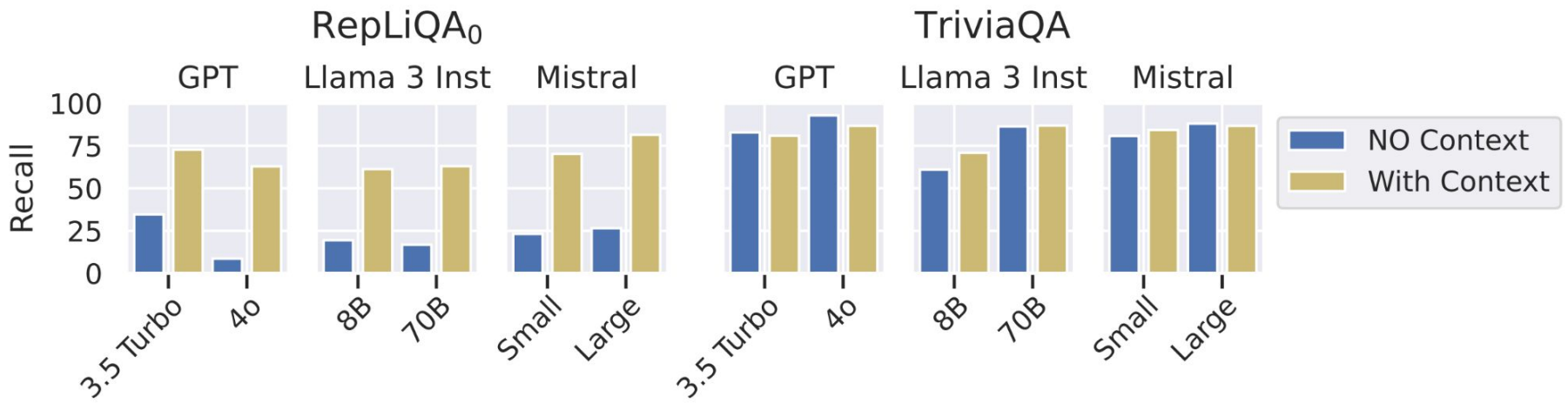


Benchmarking LLMs using RepLiQA

- All models show a **huge degradation in performance** when they cannot rely on memory

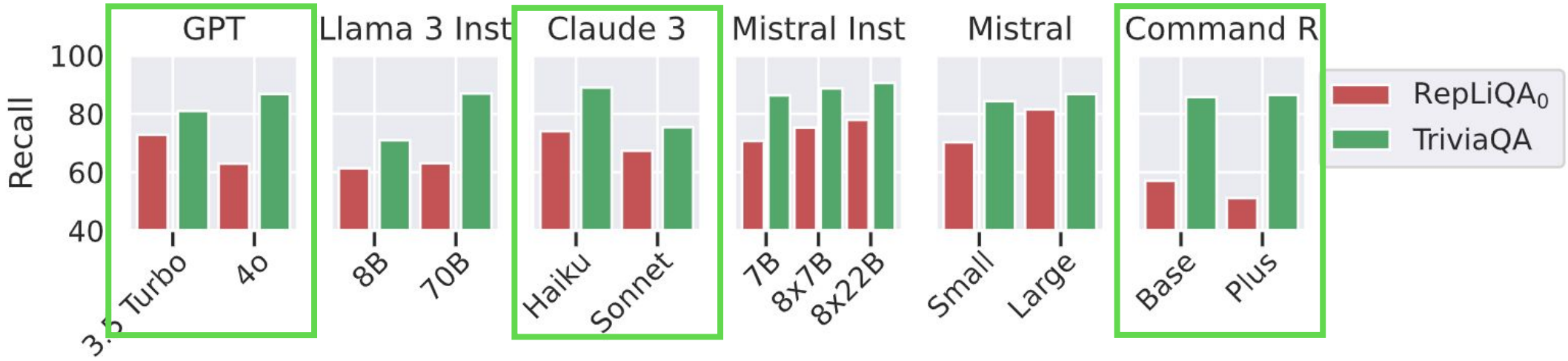
  TriviaQA covers **well-known facts**

- Models can answer TriviaQA's questions even with reference documents



However, scale alone won't help

- Generally, increasing model size improves performance on datasets such as TriviaQA
- However, **performance can drop as model grows with RepLiQA!**



SOTA LLMs seem to be rather limited in their ability to seek and use information in user-provided content.

Closing remarks

- SOTA LLMs have a rather **limited ability** to seek and use information in user-provided content.
  - ⇒ This is not captured by common benchmarks.
- **We need more benchmarks like RepLiQA.**
  - What if user-provided content contradicts model memory?
  - Should a model trust its memory or the user?
  - What if the user is an adversary?