Investigating and Building Efficient and Reliable LMs with Retrieval

Akari Asai
University of Washington
https://akariasai.github.io/
Pre-trained large language models (LLMs) have shown impressive performance across tasks, even in zero-shot.

LLMs as knowledge interface

I need to throw a dinner party for 6 people who don’t eat nuts or seafood. Can you suggest a 3-course menu?

- **Search the web for:** 3 course menu no nuts no seafood
- **Search the web for:** course menu examples

A possible 3-course menu for 6 people who don’t like nuts or seafood is:

- **Starter:** Vegetable soup with gluten-free bread. Warm and healthy dish with various veggies and herbs.
- **Main:** Roasted chicken with roasted potatoes and green beans. Classic and satisfying dish with garlic, lemon, and rosemary.
- **Dessert:** Apple crisp with vanilla ice cream. Delicious and nut-free dessert with cinnamon and gluten-free crumble.

I have sweet potato already. Can you make the main course use that?

Beginning May 7, 2025, every air traveler 18 years of age and older will need a REAL ID–compliant driver’s license or ID card [1][2], or another acceptable form of identification [3], to fly within the United States and enter certain federal facilities [4].

1. Flying with a REAL ID
   https://www.tsa.gov/real-id

2. DHS Announces Extension of REAL ID Full Enforcement Deadline ...
   https://www.dhs.gov/news/2022/12/05/dhs-announces-extension-real-id-full-enforce...

3. REAL ID - REAL ID
   https://dps.mt.gov/divisions/oivs/real-id/Pages/default.aspx

4. REAL ID
   https://www.michigan.gov/sos/license-id/real-id
Hallucinations

Their outputs often include **factual errors**.

Akari Asai is a researcher in the field of natural language processing (NLP) and machine learning. She is currently a Research Scientist at Google AI in Tokyo, Japan, and an Adjunct Professor at the University of Tokyo.

Asai has made several significant contributions to the field of NLP, particularly in the areas of machine learning for language processing, dialogue systems, and information retrieval. She has also worked on developing machine learning models that can effectively learn from limited amounts of labeled data, a problem known as few-shot learning.
Staleness

Their world knowledge gets obsolete quickly.

GPT-3

Who is the Prime Minister of Japan?

The Prime Minister of Japan is Yoshihide Suga.

Chat GPT

Who is the prime minister of Japan?

As of my knowledge cutoff date of September 2021, the Prime Minister of Japan was Yoshihide Suga. However, please note that as an AI language model, I am not constantly updated with real-time information, so the current Prime Minister of Japan may have changed since then.

Yoshihide Suga (PM: 2020-2021)  Fumio Kishida (PM: 2021-Present)
Large numbers of parameters are required to memorize.

Memorizing everything in parameters

Relying on LLMs’ **parametric knowledge** learned during pre-training results in staleness, scale, and hallucinations.

Who is the Prime minister of Japan? → LM → Yoshihide Suga

... *Yoshihide Suga* is the current prime minister of Japan as of 2021...
Can we leverage external “non-parametric” memories?

Relying on LLMs’ **parametric knowledge** learned during pre-training results in staleness, scale, and hallucinations.
Retrieval-based LM

(Guu et al., 2020; Lewis et al., 2020; Khandelwal et al., 2020; Izacard et al., 2022)

More details are in our ACL 2023 Tutorial: https://acl2023-retrieval-lm.github.io/
Develop **efficient** and **reliable** NLP systems with **retrieval** for information access for all

- Analyze the limitations of LLMs Memorizations
  (Mallen*, Asai* et al., 2023)

- Develop general-purpose **retrieval** systems
  (Asai et al., 2023)

- Improving **generation** with better control
  (Asai et al., 2022; Asai et al., 2023*)
When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories

Alex Mallen*, Akari Asai*, Victor Zhong, Rajarshi Das, Daniel Khashabi, Hannaneh Hajishirzi

* = core contributors
When do LLMs actually memorize?

Q: Who is Barack Obama?

A: Barack Obama is the 44th US President from 2009 to 2017.

Q: What is Kathy Saltzman’s occupation?

A: An actress

GPT-3
Investigating effectiveness of parametric and non-parametric memories

**RQ1:** How much knowledge is memorized and what factors affect memorizations?

**RQ2:** Can non-parametric memories alleviate shortcomings of parametric memories?

**RQ 3:** Can we effectively combine non-parametric and parametric memories?
RQ1: How much knowledge is memorized and what factors affect memorizations?

RQ2: Can non-parametric memories alleviate shortcomings of parametric memories?

RQ 3: Can we effectively combine non-parametric and parametric memories?
Investigating effectiveness of parametric and non-parametric memories

**RQ1:** How much knowledge is memorized and What factors affect memorizations?

**RQ2:** Can non-parametric memories alleviate shortcomings of parametric memories?

**RQ 3:** Can we effectively combine non-parametric and parametric memories?
Focus and task

Focus: Factual knowledge

Task: Open-domain QA

Q: What is the occupation of Kathy Saltzman?
A: politician

(Kathy Saltzman, occupation, Politician)
Dimensions of analysis

- Subject entity popularity (= Wikipedia page view)

(Kathy Saltzman, occupation, Politician) < (Barack Obama, occupation, Politician)
Dimensions of analysis

• Subject entity popularity (= Wikipedia page view)

(Kathy Saltzman, occupation, Politician)

< (Barack Obama, occupation, Politician)

• Relationship types

(Lionel Messi, father, Jorge Messi) < (Lionel Messi, sports, soccer)
Dataset: PopQA

1. Sampling factual knowledge

(Louisiana, capital of, Baton Rouge)

(Kathy Saltzman, occupation, politician)

Knowledge triples from Wikidata

2. Converting triples to questions

Q: What is the capital of Louisiana?
A: Baton Rouge

Q: What is Kathy Saltzman’s occupation?
A: Politician

3. Collect popularity

Wikipedia pageview
Investigating effectiveness of paramedic and non-parametric memories

**RQ1:** How much knowledge is memorized and what factors affect memorizations?

**RQ2:** Can non-parametric memories alleviate shortcomings of parametric memories?

**RQ 3:** Can we effectively combine non-parametric and parametric memories?
Does popularity predict factual knowledge?

Strong correlations across 10 LMs, with a greater correlation in larger LMs.
Does this hold across relationship types?

Positive correlations hold across different relationship types.
Performance increases as model scale — does scaling solve everything?
Scaling may not help!

Scaling helps to memorize popular factual knowledge
Scaling may not help!

Scaling helps to memorize **popular factual knowledge**

Almost zero improvements on **less popular factual knowledge**
RQ1: How much knowledge is memorized and what factors affect memorizations?

RQ2: Can non-parametric memories alleviate shortcomings of parametric memories?

RQ 3: Can we effectively combine non-parametric and parametric memories?
Retrieval-in-context prompting

Who was the director of The White Suit?

In 1999 "The White Suit" an auteur film by Ristovski (director, writer, lead actor, and producer) was at the Cannes Film Festival in...

Q: Who was the director of The White Suit?
A:
Does retrieval really help?

A simple retrieval-in-context gives large gains across LMs
Does retrieval really help?

GPT-1.3B with retrieval can outperform GPT-3

Accuracy (%)
When does retrieval help?

Non-parametric knowledge helps in **less popular entities**.
When does retrieval help?

Non-parametric knowledge helps in **less popular entities**.

Parametric memories only models outperforms in **popular entities**.
Why does retrieval sometimes hurt?

Retrieval causes LMs to counterfactually fail due to low retrieval recall.

Recall @ 1

- Correct answer
- Wrong answer

Retrieval helps

Retrieval hurts
Investigating effectiveness of parametric and non-parametric memories

RQ1: How much knowledge is memorized and What factors affect memorizations?

RQ2: Can non-parametric memories alleviate shortcomings of parametric memories?

RQ 3: Can we effectively combine non-parametric and parametric memories?
Adaptive retrieval: retrieves when necessary

Input

Popularity > threshold?

Yes → Standard LM

No → Retrieval-augmented LM

Not memorized in parameters ⇔ use retrieval
Memorized in parameters ⇔ don’t use retrieval

What is Kathy Saltzman’s occupation?

What is the capital of Louisiana?

Accuracy

unassisted LM
retrieval-augmented

Popularity

$10^1$ $10^2$ $10^3$ $10^4$ $10^5$
Adaptive retrieval improves performance

Larger performance gains with more competitive LMs

Accuracy (%)

- vanilla
- adaptive
- always

GPT-6B
GPT-20B
GPT-3
Adaptive retrieval improves efficiency

Adaptive retrieval can also reduce inference-time costs

Latency (sec) / query

API cost ($) / query

GPT-j-6B
GPT-Neo-20B
GPT-3
Task-aware retrieval with instructions

Akari Asai, Timo Schick, Patrick Lewis, Xilun Chen, Gautier Izacard, Sebastian Riedel, Hannaneh Hajishirzi, Wen-tau Yih
Retrieval struggles in OOD

DPR underperforms BM25 in **zero-shot** retrieval.

---

Specialized retrievers v.s. a generalizable retriever

Asai et al. “Task-aware retrieval with instructions”. In ACL 2023 Findings.
An instruction-following task-aware retriever

$q$: Implementing batch normalization in Python

$t_1$: I want to find python code

def batchnorm_backward(dout, cache):
    N = dout.shape[0]
    x_norm, x_centered, std, gamma = cache

$I_2$: Retrieve related questions

I am developing batch normalization from scratch. How can I do that?

$t_2$: Please find answers

Answer:
You can just `import torch.nn.BatchNorm2d`

Better generalizations & instruction-following

Instruction-tuning for retrieval shows effectiveness in zero-shot & cross-task

<table>
<thead>
<tr>
<th>Method</th>
<th>BEIR</th>
<th>Cross-Task</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contriever</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Contriever+CE</td>
<td>45</td>
<td>45</td>
</tr>
<tr>
<td>TART Dual</td>
<td>30</td>
<td>50</td>
</tr>
<tr>
<td>TART Full</td>
<td>15</td>
<td>60</td>
</tr>
</tbody>
</table>

NDCG@10
Better generalizations & instruction-following

Instruction-tuning for retrieval shows effectiveness in zero-shot & cross-task

NDCG@10

- Blue: Contriever
- Green: Contriever+CE
- Purple: TART Dual
- Orange: TART Full

BEIR

Cross-Task
Develop efficient and trustworthy NLP systems with retrieval for information access for all

- Analyze the limitations of LLMs Memorizations
  (Mallen*, Asai* et al., 2023)

- Develop general-purpose retrieval systems
  (Asai et al., 2023)

- Improving generation with better control
  (Asai et al., 2022; Asai et al., 2023*)

Contact: Akari Asai
akari@cs.washington.edu
https://akariasai.github.io/


Asai et al. "Task-aware retrieval with instructions”. In ACL 2023 Findings.