Adaptive and trustworthy NLP with Retrieval for information access for everyone

Akari Asai

W

UNIVERSITY of WASHINGTON
Pre-trained large language models (LLMs) have shown impressive performance across tasks, even in zero-shot.

Large language models 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?

- **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/travel/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.mn.gov/divisions/civs/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.

Asai has published numerous research papers in top-tier NLP and machine learning conferences and journals, including ACL, EMNLP, and NeurIPS. She has received several awards for her research, including the Best Long Paper Award at EMNLP 2018, the Best Paper Award at SIGDIAL 2014, and the Google PhD Fellowship in NLP in 2014.

Overall, Akari Asai is a highly respected and accomplished researcher in the field of NLP, known for her innovative work on machine learning models for language processing and dialogue systems.
Staleness

Their world knowledge gets obsolete quickly.

Yoshihide Suga (PM: 2020-2021)
Fumio Kishida (PM: 2021-Present)

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.
Extremely large numbers of parameters to memorize everything.

Limitation of memorizing everything in parameters

Relying on LLMs’ **parametric knowledge** learned from pretraining results in staleness, scale, and hallucinations.

Who is the Prime minister of Japan?

... **Yoshihide Suga** is the current prime minister of Japan ...

Yoshihide Suga
Memorizing everything in parameters

Can we leverage external “non-parametric” memories?

Relying on LLMs’ **parametric knowledge** results in staleness, scale, and hallucinations.

Who is the Prime minister of Japan?

... **Fumio Kishida** replaced Yoshihide Suga on October 4, 2021

Fumio Kishida
Develop efficient and trustworthy NLP systems with retrieval for information access for everyone.

Non-parametric memories
(Retrieved web documents)

Parametric memories
(LLMs)
Who is the Prime minister of Japan?

... Fumio Kishida replaced Yoshihide Suga on October 4, 2021

Fumio Kishida

Develop **efficient** and **trustworthy** NLP systems with **retrieval** for information access for everyone

- **Analyze the limitations of LLMs Memorizations**
  (Mallen*, Asai* et al., 2022)

- **Develop general-purpose retrieval systems**
  (Asai et al., 2022)
  (Lin, Asai et al., 2023)

- **Improving generation with better control**
  (Asai et al., 2022)
Analyze the limitations of LLMs Memorizations
(Mallen*, Asai* et al., 2022)

Why do we need both parametric and non-parametric memories?

How can we develop general-purpose retrieval systems?

How can we make a language model to pay attention to relevant evidence?

Develop **efficient** and **trustworthy** NLP systems with **retrieval** for information access for everyone

- **Analyze the limitations of LLMs Memorizations** (Mallen*, Asai* et al., 2022)
- **Develop general-purpose retrieval systems** (Asai et al., 2022) (Lin, Asai et al., 2023)
- **Improving generation with better control** (Asai et al., 2022)

Apply retrieval-augmented LMs for low-resource domains (e.g., multilingual)
Investigating effectiveness of parametric and non-parametric memories
How much knowledge is *memorized*?

How much knowledge is *memorized*?

How much knowledge is memorized?

Roberts et al. “How much knowledge can you pack into the parameters of a language model?”. In EMNLP. 2020.
How much knowledge is *memorized*?

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?

Focus and Task

Focus: Factual knowledge

(Task: Open-domain QA

Q: What is the occupation of Kathy Saltzman?
A: 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)
  - (Lionel Messi, father, Jorge Messi)
  - (Lionel Messi, sports, soccer)

Popularity biases in existing QA datasets

Many existing QA datasets (e.g., Natural Questions; Kwiatkowski et al., 2019) are often bias towards popular entities.

Dataset: PopQA

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?

RQ3: Can we effectively combine non-parametric and parametric memories?

Memorization depends on popularity

Strong correlation between popularity and accuracy

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

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?

Retrieval-augmented LM

Retrieve-then-read (Retriever + LM)

Who is the Prime minister of Japan? + ... Fumio Kishida replaced Yoshihide Suga on October 4, 2021 → Fumio Kishida

Results of retrieval-augmented LM

Non-parametric retrievers gives Large improvements

Results of retrieval-augmented LM

1.5B GPT+retriever is better than GPT-3 (003; 175B?)

When retrieval helps?

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

When retrieval helps?

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

Parametric memories only models outperforms in **popular entities**.

Qualitative analysis on the errors

<table>
<thead>
<tr>
<th>Question (answer)</th>
<th>LM</th>
<th>Retrieval-LM</th>
<th>Retrieval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Who was the director of Black? (Sanjay Leela Bhansali)</td>
<td>The director of Black is Sanjay Leela Bhansali. ✓</td>
<td>Noel Black ✗</td>
<td>Noel Black (June 30, 1937 – July 5, 2014) was an American film and television director, screenwriter, and producer. Black was born in Chicago, Illinois. He won awards at the 1966 ... and subsequently concentrated on directing for television, occasionally directing films such as &quot;Private School&quot;.</td>
</tr>
</tbody>
</table>

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

Adaptively combine parametric and non-parametric memories to get the best of the both worlds.

- **Popularity > threshold?**
  - Yes: Standard LM
  - No: Retrieval-augmented LM

- **Not memorized in parameters** ⇔ use retrieval

**What is Kathy Saltzman's occupation?**

**Memorized in parameters** ⇔ don't use retrieval

**What is the capital of Louisiana?**

Adaptive retrieval for performance

Adaptive Retrieval improves performance across LMs

Adaptive retrieval for efficiency

Adaptive Retrieval improves performance across LMs

Latency (sec) / query

<table>
<thead>
<tr>
<th>Model</th>
<th>vanilla</th>
<th>Adaptive</th>
<th>Always</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPT-j-6B</td>
<td>2</td>
<td>6</td>
<td>8</td>
</tr>
<tr>
<td>GPT-Neo-20B</td>
<td>4</td>
<td>5</td>
<td>7</td>
</tr>
</tbody>
</table>

API cost ($) / query

<table>
<thead>
<tr>
<th>Model</th>
<th>vanilla</th>
<th>Adaptive</th>
<th>Always</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPT-3</td>
<td>1</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>

Takeaways

1. Scaling may not help LLMs memorize long-tail knowledge.

2. Retrieval-augmentation (non-parametric memories) can complement the shortcomings of parametric memories.

3. Adaptively combine parametric and non-parametric memories gives performance & efficiency gain.

Towards adaptable general-purpose neural retrievers
Finding the right non-parametric memories is a key

Large scale of non-parametric memories (e.g., 6 millions of English Wikipedia articles)

Who is the Prime minister of Japan?

Retriever (e.g., BM 25)

... Fumio Kishida replaced Yoshihide Suga on October 4, 2021

Sanna Marin is a Finnish politician who has been serving as the PM of Finland since 2019. A member of ...

The Liberal Democratic Party is a conservative political party in Japan.
Neural pre-trained encoder-based retriever

Who is the Prime minister of Japan?

Fumio Kishida replaced Yoshihide Suga on October 4, 2021.

Sanna Marin is a Finnish politician who has been serving as the PM of Finland since 2019. A member of...

The success of neural retrievers

Neural retrievers outperform classical term-based BM25 in resource abundant settings.

Their struggles in zero-shot transfer

BM25 significantly outperforms DPR in **zero-shot** retrieval (e.g., Bio, financial news articles)

Training with instructions help an LLM to adapt to new task without training & enables better controllability

**Finetune on many tasks ("instruction-tuning")**

- **Input (Commonsense Reasoning)**
  - Here is a goal: Get a cool sleep on summer days.
  - How would you accomplish this goal?
  - OPTIONS:
    - Keep stack of pillow cases in fridge.
    - Keep stack of pillow cases in oven.
  - **Target**
    - keep stack of pillow cases in fridge

- **Input (Translation)**
  - Translate this sentence to Spanish:
    - The new office building was built in less than three months.
  - **Target**
    - El nuevo edificio de oficinas se construyó en tres meses.

**Inference on unseen task type**

- **Input (Natural Language Inference)**
  - Premise: At my age you will probably have learnt one lesson.
  - Hypothesis: It's not certain how many lessons you'll learn by your thirties.
  - Does the premise entail the hypothesis?
  - OPTIONS:
    - yes
    - it is not possible to tell
    - no
  - **FLAN Response**
    - It is not possible to tell

Building separate task-specific retrieval systems?

Training a instructable retriever

$q$: Implementing batch normalization in Python

$t_1$: I want to find python code

$t_2$: Retrieve related questions

$t_2$: Please find answers

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

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

Training a instructable retriever

\( q: \) Implementing batch normalization in Python

\( t_1: \) I want to find python code

\( t_2: \) Retrieve related questions

\( t_2: \) Please find answers

\( \text{def } \) batchnorm_backward
\( \text{(dout, cache):} \)
\( \text{N = dout.shape[0]} \)
\( x\_norm, x\_centered, \)
\( \text{std, gamma } = \text{ cache} \)

I am developing batch normalization from scratch

Answer:
You can just `import torch.nn.BatchNorm2d`
Instructions scheme for retrieval tasks

Our diverse instructions satisfy three key components to define a retrieval task: **Intent**, **Domain** and **Unit**.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Instruction</th>
</tr>
</thead>
<tbody>
<tr>
<td>NQ</td>
<td>Retrieve a <em>Wikipedia paragraph</em> that answers this question.</td>
</tr>
<tr>
<td>QReCC</td>
<td>Find a <em>dialogue response</em> from dialogue history to answer the user’s question.</td>
</tr>
<tr>
<td>Arguana</td>
<td>Retrieve a <em>paragraph</em> from an argument website that argues against the following argument.</td>
</tr>
<tr>
<td>SciFact</td>
<td>Find a <em>sentence</em> from a <em>scientific paper</em> to check if the statement is correct or not.</td>
</tr>
<tr>
<td>MultiLexSum</td>
<td>I want to find the <em>one-sentence summary</em> of this legal case.</td>
</tr>
</tbody>
</table>

**BERRI**: large-scale dataset with Instructions

Retrieve a Wikipedia article that answers the following question: Who is the Prime minister of Japan?

Sanna Marin is a Finnish politician who has been serving as the PM of Finland since 2019. A member of ...

... Fumio Kishida replaced Yoshihide Suga on October 4, 2021

Approximate nearest neighbor search

(Base LM: BERT-base)

Retrieve a Wikipedia article that answers the following question: Who is the Prime minister of Japan?

Fumio Kishida replaced Yoshihide Suga on October 4, 2021.

Retrieve a Wikipedia article that answers the following question: Who is the Prime minister of Japan?

Sanna Marin is a Finnish politician who has been serving as the PM of Finland since 2019. A member of ...

(Base LM: T0 / FLAN T5)

Training

\[ \mathcal{L} = - \log \frac{e^s(q,d^+,t)}{\sum_{d \in B} e^s(q,d,t)} \]

\[ \mathcal{L} = - \sum_{d \in d^+} \log s(t, q, d) - \sum_{d \in d^-} \log(1 - s(t, q, d)) \]

Zero-shot evaluation

**BEIR** (Thakur et al., 2021) and **LOTTE** (Khattab et al., 2022)

Results

Instruction-tuned retrievers help in zero-shot & open domain

Effect of instructions

Instruction-tuned retrievers help in zero-shot & open domain

Takeaways

1. Shift IR paradigm from multiple task-specific models to a single unified model users can *instruct*.

2. Construct a large-scale dataset with instructions for instruction tuning for retrieval.

3. Build TART, a single task-aware retrieval system and achieves strong zero-shot and open-domain retrieval performance.

Trustworthy generation based on reliable evidence
Hallucinations even with the right links

There are a lot of things going on in Ohio, but one of the most serious ones is the toxic train derailment that happened near East Palestine on 3 February. A train carrying hazardous chemicals, including sulfuric acid, hydrochloric acid, sodium hydroxide and potassium hydroxide, derailed and caught fire, causing a massive chemical leak that contaminated the nearby waterways and soil. The residents of East Palestine have filed several class action lawsuits against the train company, Norfolk Southern, alleging negligence and seeking damages and medical monitoring. Some have compared the situation to a “Chernobyl 2.0” because of the long-term environmental and health effects.
Knowledge conflicts in memories

Which country won the most medals in winter olympics?

Non-parametric Knowledge (Documents retrieved at inference time)

- Passage 1: Norway set the record for most total medals at a single Winter Olympics with 39, surpassing the...
- Passage 2: Norway was the most successful nation at the games with 39 total medals, setting a new record for the most medals won by a country at a single Winter Olympics.
- Passage 3: With 36 total medals, Germany set a record for most total medals at a Winter Olympics...

Parametric Knowledge (Facts memorized during training)

The U.S. team had a historic Winter Games, winning an unprecedented 37 medals.

Inconsistencies between training data and up-to-date articles

Knowledge conflicts in memories

Spurious correlations during training

Standard training can amplify the reliance on spurious evidence.

How many countries does India border?

Passage with evidence:
- India shares the border with **seven** countries.
- India is the **seventh** largest countries.

Passage without evidence:
- **y = G(x, P)**
Spurious correlations during training

Standard training can amplify the reliance on spurious evidence.

How many countries does India border?

Passage with evidence:
India shares the border with seven countries.
India is the seventh largest countries.

Passage without evidence:

\[ y = G(x, P) \]

seven \( \hat{y} \)
seven \( y \)
Evidentiality of non-parametric memories

Generator should consider **evidentiality** (Lee et al., 2021) and focuses on evidentiality-positive non-parametric memories.

India shares the border with **seven** countries.

**Evidentiality-positive** passages include sufficient evidence to provide right outputs.

India is the **seventh** largest countries.

**Evidentiality-negative** passages don’t, even though they can happen to include answers.

Learn to predict evidentiality and generate

How many countries does India border?

Passage with evidence
- India shares the border with seven countries.
- India is the seventh largest countries.

Passage without evidence

$L = L_{gen} + \lambda L_{class}$
Designing evidentiality-guided generator

How many countries does India border?

Passage with evidence
- India shares the border with **seven** countries.
- India is the **seventh** largest countries.

Passage without evidence

Formal expression:

\[ \mathcal{L} = \mathcal{L}_{\text{gen}} + \lambda \mathcal{L}_{\text{class}} \]
Standard encoder-decoder LM
Designing evidentiality-guided generator

Encoder

\( P_1 \)

\( P_2 \)

Answer Decoder

\( y \)

Evidentiality Predictor

Positive

Negative
Designing evidentiality-guided generator

Encoder

P₁

P₂

Answer Decoder

y

Lₖₐₜ

Positive

Negative

Lₘₜₙ
How many countries does India border?

Passage with evidence:
- India shares the border with seven countries.

Passage without evidence:
- India is the seventh largest countries.

Collecting silver evidentiality labels

$L = L_{gen} + \lambda L_{class}$
How many countries does India border?

- Passage with evidence:
  - India shares the border with seven countries.
  - India is the seventh largest countries. It has their

LM-based data generation
Evaluation: QA, Fact checking and dialogue

Open-domain QA

How many countries does India border?
Seven

Fact Checking

India shares the border with seven countries.
True

Knowledge-grounded Dialogue

> user: You know India shares border with many countries. Which country have you been to?
> AI: Yes! I actually been to Nepal!
Strong performance across three tasks

Takeaways

1. Models don’t effectively use non-parametric memories

2. Introduce evidentiality-guided generator to make LM to focus on reliable information

3. Show effectiveness on diverse knowledge intensive tasks including QA, dialogue, fact verification

Summary
Develop efficient and trustworthy NLP systems with retrieval for information access for everyone.

- Analyze the limitations of LLMs Memorizations
- Develop general-purpose retrieval systems
- Improving generation with better control

- Using non-parametric memories can complement LMs
- Diverse training strategies help to build generalizable retrievers
- Enforcing models to pay more attentions to relevant evidence further improves performance.
Develop **efficient** and **trustworthy** NLP systems with retrieval for information access for everyone.

(Near) Future research of retrieval-augmented LMs

“Adaptive” retrieval-augmented LM

controllability and verifiability

More diverse applications
(Near) Future research of retrieval-augmented LMs

“Adaptive” retrieval-augmented LM
controllability and verifiability
More diverse applications

User
What kind of animal is Scooby from Scooby Doo?
GopherCite
A Great Dane dog.

Wikipedia Page: Scooby-Doo
This Saturday-morning cartoon series featured teenagers Fred Jones, Daphne Blake, Velma Dinkley, and Shaggy Rogers, and their talking Great Dane named Scooby-Doo.

(Near) Future research of retrieval-augmented LMs

“Adaptive” retrieval-augmented LM

Controllability and verifiability

More diverse applications

Develop **efficient** and **trustworthy** NLP systems with **retrieval** for information access for everyone

- Analyze the limitations of LLMs Memorizations (Mallen*, Asai* et al., 2022)
- Develop general-purpose **retrieval** systems (Asai et al., 2022) (Lin, Asai et al., 2023)
- Improving **generation** with better control (Asai et al., 2022)

Akari Asai | Twitter: @AkariAsai | Email: akari@cs.washington.edu
https://akariasai.github.io/