

RAG & Memory

CS6960 MultiModal LLM Agents

Kenneth Marino

Announcements

- HW1 due Friday
 - See Syeda's note in announcements / read the new instructions
 - Can extend deadline if people are still having issues
- Projects
 - Fill in form in Piazza with ideas/area of interest
 - In-class activity with people in your interest area
 - CHPC/Useful resources doc on Canvas

Recommended Readings

- Agent Courses/Tutorials

- <https://www.shuyanzhou.com/teaching/25fall-590/25fall-590.html>
- <https://rdi.berkeley.edu/adv-llm-agents/sp25>
- <https://agenticai-learning.org/sp25>

- This week's papers

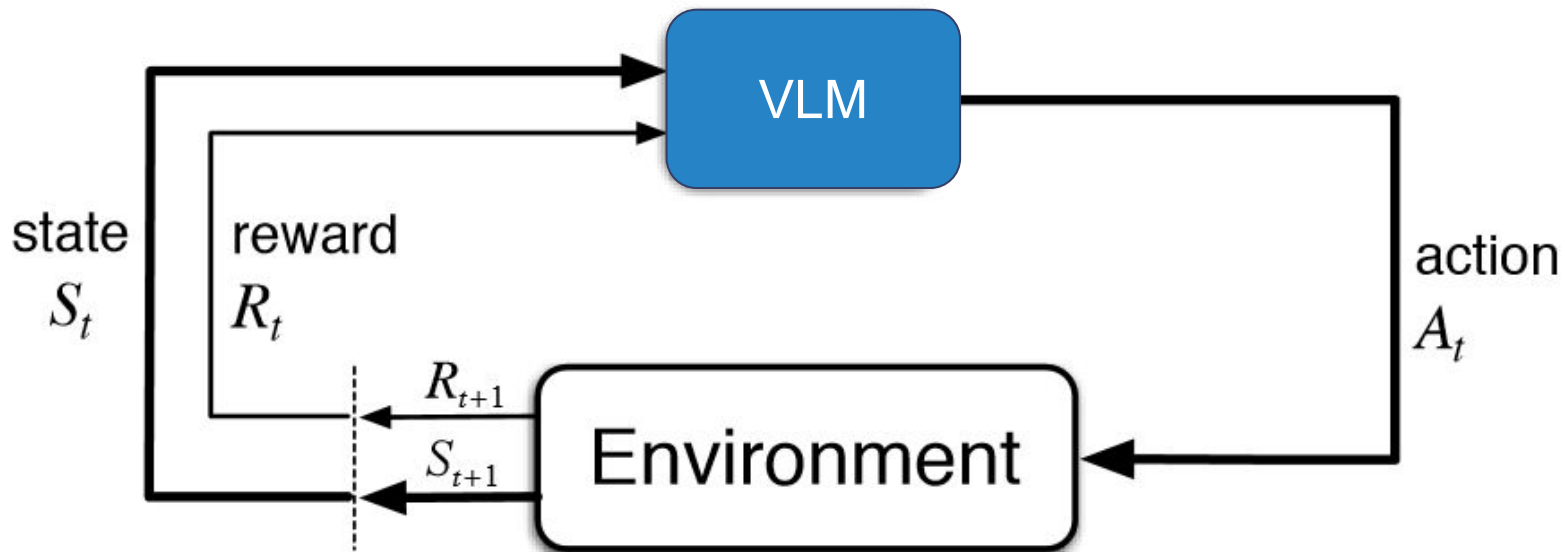
- [Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks](#)
- [REALM: Retrieval-Augmented Language Model Pre-Training](#)
- [Dense Passage Retrieval for Open-Domain Question Answering](#)
- [GNN-RAG: Graph Neural Retrieval for Large Language Model Reasoning](#)
- [MemGPT: Towards LLMs as Operating Systems](#)

Lectures Going Forward

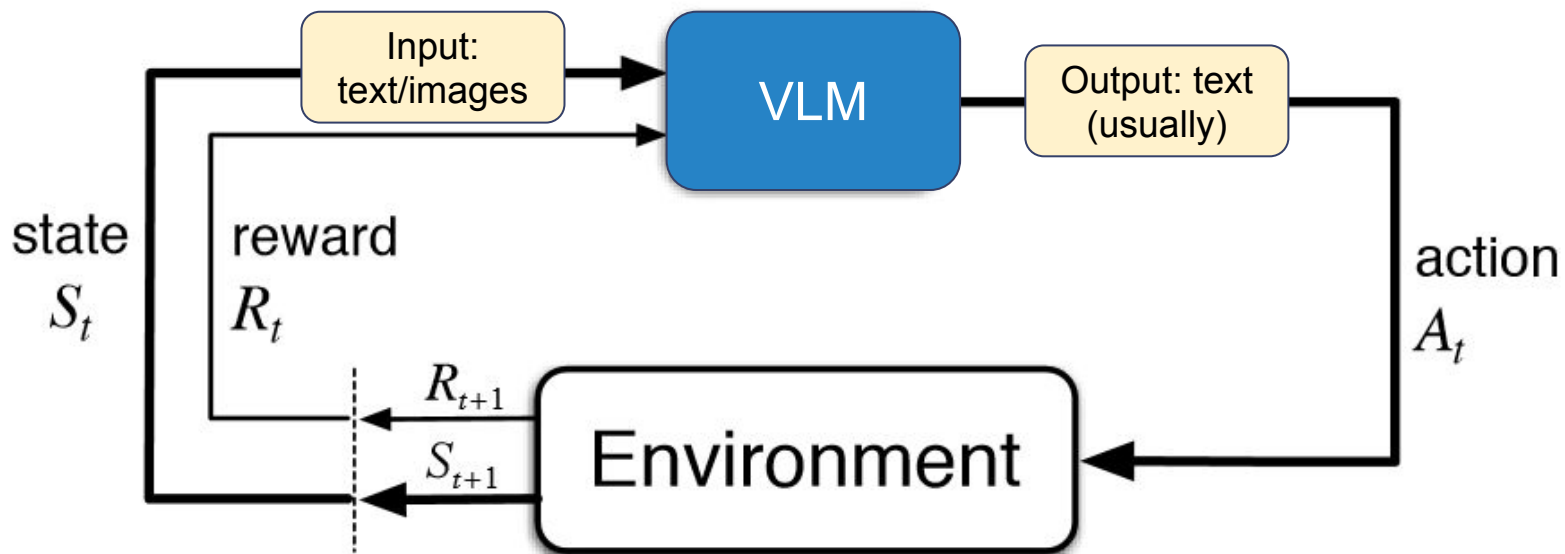
- Meant to be fairly short / high-level
 - More an introduction/framing of the topic than a deep dive
 - Maybe 20-40 mins depending on how much I have to say about the topic
- Imperative that you read papers and participate in class discussions to get full breadth of topic

Any Questions

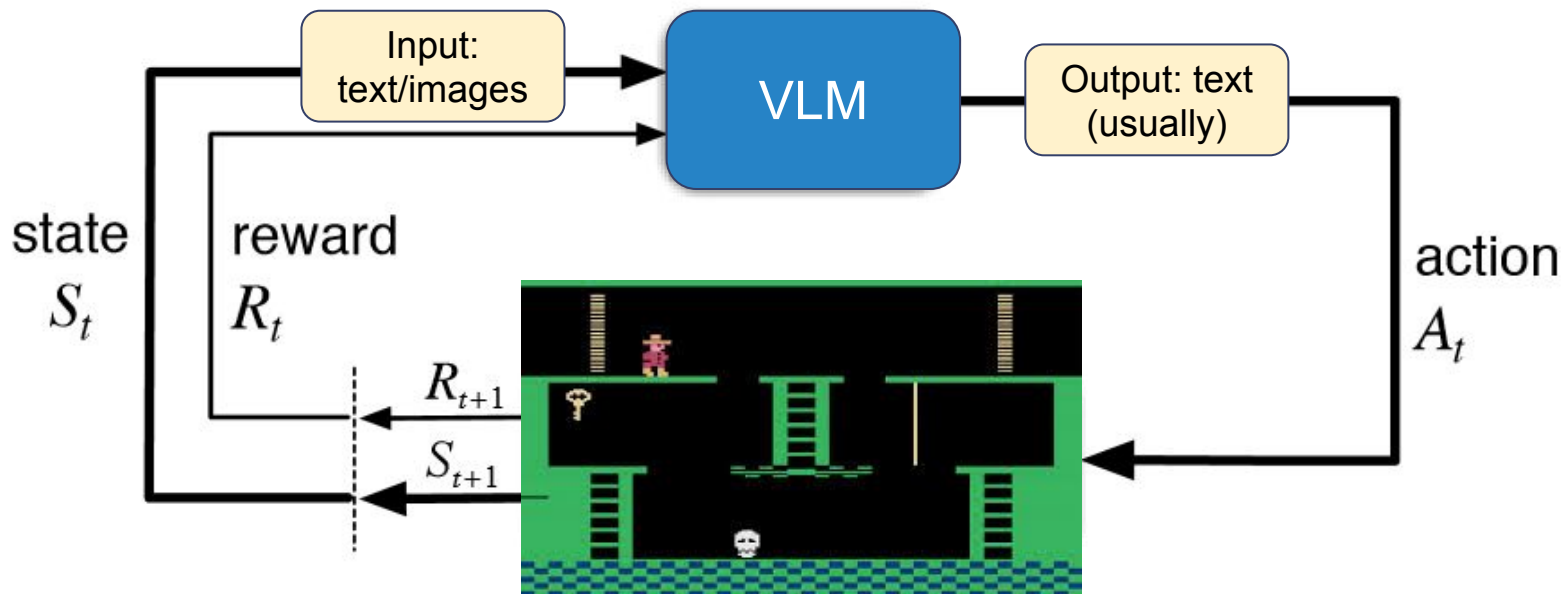
What does VLM agent actually look like?



What does VLM agent actually look like?



What does VLM agent actually look like?



Other “stuff”

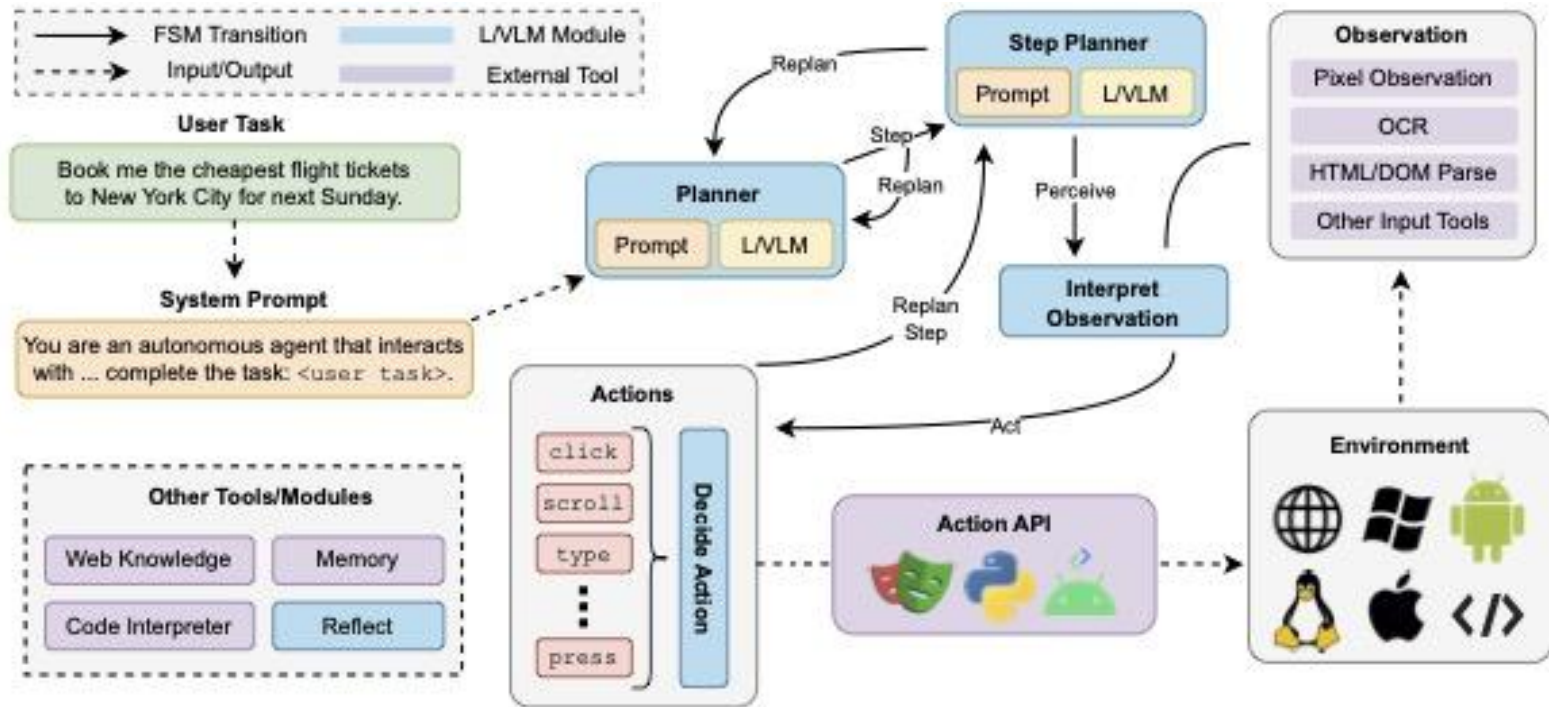
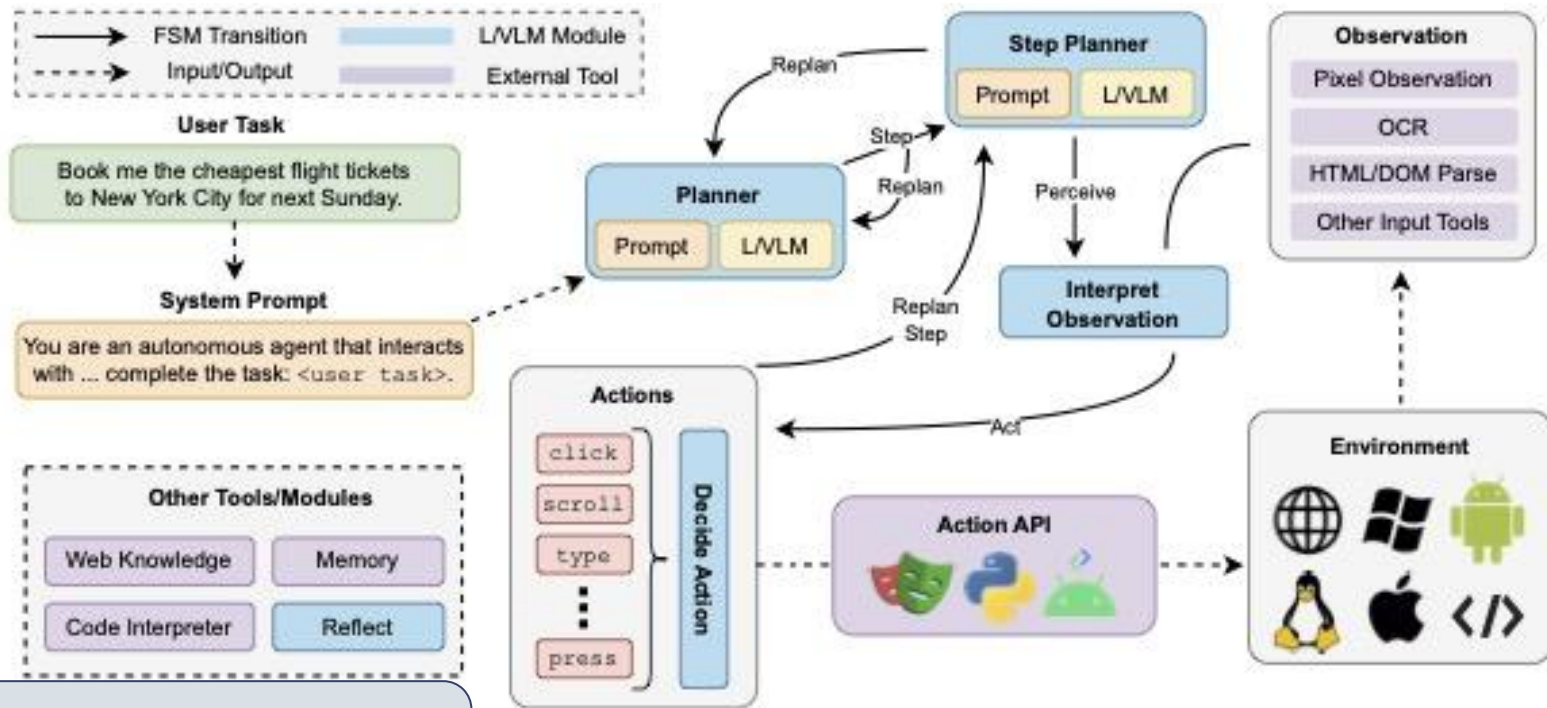


Image credit: Farhan Ishmam

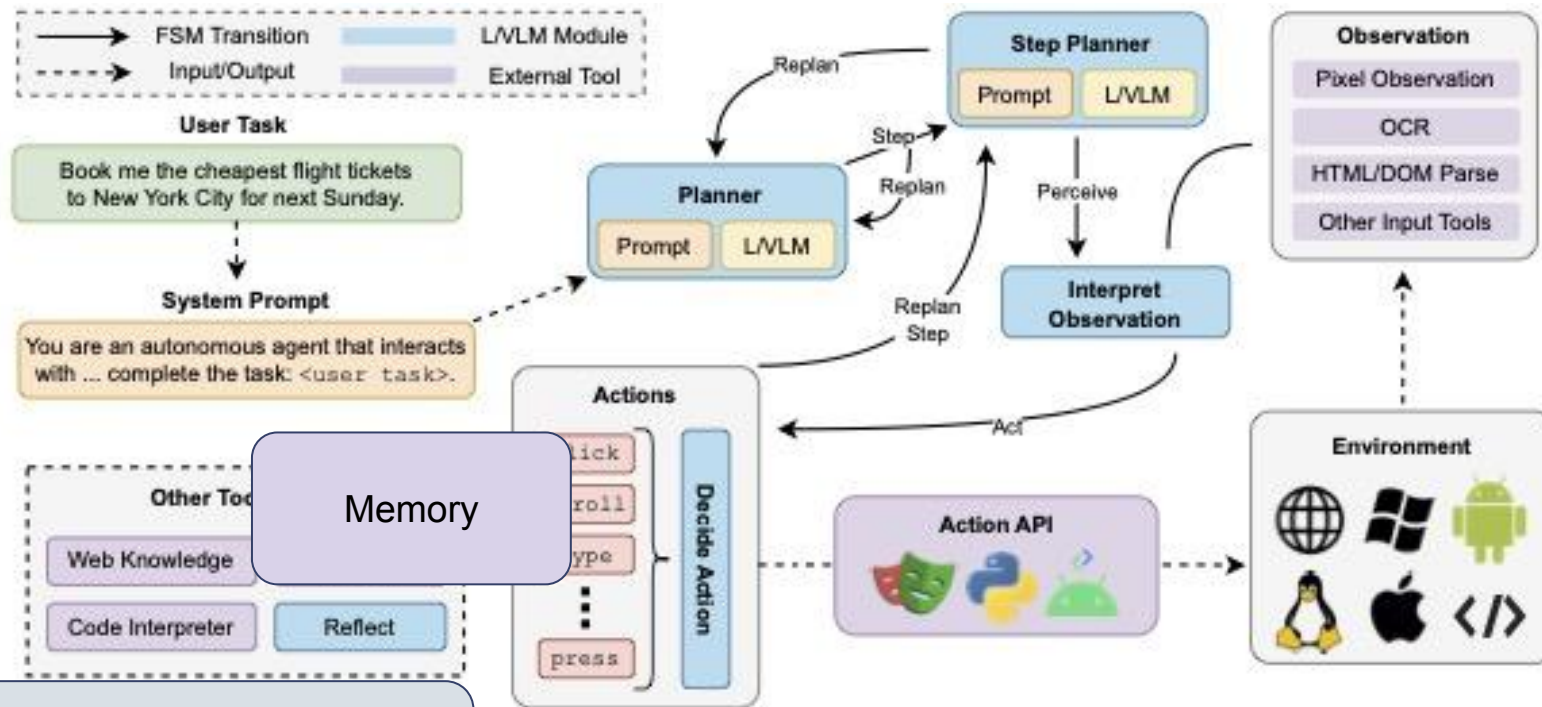
Other “stuff”



Tools VLM can invoke that do different things

Image credit: Farhan Ishmam

Other “stuff”



Tools VLM can invoke that do different things

Recall: Markov Assumption

- Markov Assumption
 - For environment - only your current state matters, not how you got there
 - For policy - you only need to look at your current state
- Fully Observable
 - Everything about the environment is full observed and captured in your state s

Example: Computer Use

The image displays a web browser window showing a product page on 'One Stop Market' and a terminal window displaying the output of a Python script using GPT-4 for vision analysis.

Browser Window (Left):

- URL: metis.lli.cs.cmu.edu:7770
- Page Title: One Stop Market
- Search bar: Search entire store here...
- Navigation: My Account, My Wish List, Sign Out
- Categories: Personal Care, Sports & Outdoors, Clothing, Shoes & Jewelry, Home & Kitchen, Office Products, Tools & Home Improvement, Household, Patio, Lawn & Garden, Electronics, Cell Phones & Accessories, Video Games, Grocery & Gourmet Food
- Product Grid:

 - Red Gingerbread House Kit**: Pack, 17 oz., Pack of 2, Total. Price: \$14.47. 1 Review.
 - V8 +Energy, Healthy Energy Drink**: Steady Energy from Black and Green Tea, Pomegranate Blueberry, 8 Ounce Can, Pack of 24. Price: \$19.36. 4 Reviews.
 - Elmwood Inn Fine Tea, Orange Vanilla Caffeine-free Fruit Infusion, 16-Ounce Pouch**. Price: \$23.50. 4 Reviews.
 - Belle Of The Ball Princess Sprinkle Mix | Wedding Colorful Sprinkles | Cake Cupcake Cookie Sprinkles | Ice Cream Candy Sprinkles | Yellow Gold Red Royal Red Rose Icing Flowers Decorating Sprinkles, 8OZ**. Price: \$15.62. 12 Reviews.
 - So Delicious Dairy Free CocoWhip**: Light, Vegan, Non-GMO Project Verified, 9 oz. Tub. Price: \$15.62. 12 Reviews.

Terminal Window (Right):

```
visualwebarena_mem - node - python run_eval.py --instruction_path
Predicted Next State: The next state will show the comment section with a
led "Interested" containing the text "Do you have a USB-C cable?".
Action: In summary, the next action I will perform is '''type [20] [Interes
owed by '''type [21] [Do you have a USB-C cable?]' and then '''click [22]
Running gpt-4-vision-preview...
Plan: The image of the Seagate 1tb external hard drive shows a cable, but I
an if the cable is USB-C or another type. Since the objective requires con
a USB-C cable, and the previous actions show that I have already asked the
cy have a USB-C cable. I will wait for the seller's response before proced
comment has been approved, no further action is required at this moment.
Summary: The task of inquiring about the USB-C cable has been completed, an
for the seller's response.
Predicted Next State: The next state will remain the same until the seller
the comment.
Action: In summary, the next action I will perform is '''stop [1]'''
2024-03-20 02:47:48,186 - INFO - [Result] (PASS) config_files/test_classifi
Saving jsons...
Iteration 3/5
2024-03-20 02:47:48,186 - INFO - [Config file]: config_files/test_classific
2024-03-20 02:47:48,186 - INFO - [Intent]: What is the email of the seller
palette on this page?
Running gpt-4-vision-preview...
Plan: The objective is to find the email of the seller of the red palette.
STARLIGHT Palette listed for $10 seems to be the red palette mentioned. The
ID [24] corresponds to this palette. To get the seller's email, I should cl
image or the associated link with ID [25]. By doing so, I expect to be take
led view where the seller's contact information is typically provided.
Summary: Clicking on the image or link of the Lancoem STARLIGHT Palette will
me to a detailed page where the seller's email can be found, thus progress
the objective.
Predicted Next State: After clicking on the listing for the "Lancoem STARLI
Brand How" the next state will display the detailed view of the listing,
images of the palette, a description, options to contact the seller or share
e and a section for comments. There will also be information about the sell
& their name and email address for contact purposes.
Action: In summary, the next action I will perform is '''click [25]'''
Running gpt-4-vision-preview...
Plan: The objective is to find the email of the seller of the red palette.
on this page. I have already clicked on the listing for the Lancoem STARLI
which is the red palette mentioned in the objective. The seller's contact i
available on the page, including the email address. Since obtaining the em
was the objective, I can now complete the task by issuing the stop action w
1 address as the answer.
Summary: The action retrieves the seller's email from the listing page of t
te, fulfilling the task's objective to obtain this contact information.
Predicted Next State: The episode will end after the stop action is issued
reaction provided will be returned to the user.
Action: In summary, the next action I will perform is '''stop [sofia_kumar@
Saving jsons...
Iteration 4/5
Input Image: https://images.pexels.com/photos/2547893/pexels-photo-2547893
2024-03-20 02:48:25,759 - INFO - [Config file]: config_files/test_shopping
2024-03-20 02:48:25,759 - INFO - [Intent]: Find me powder to make the batter
the same as the picture.
[1]
```

Example: Computer Use

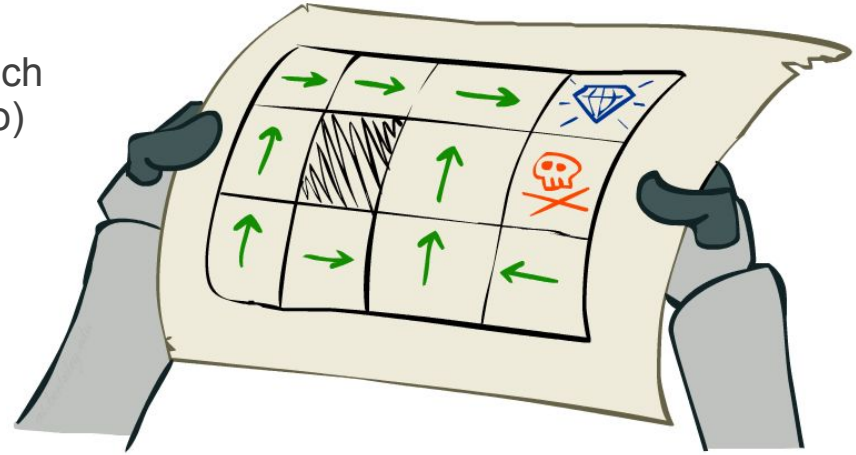
The screenshot displays a web browser window with the URL `metis.lli.cs.cmu.edu:7770` and the page title "One Stop Market". The page shows a grid of products. A terminal window is overlaid on the right side, showing the following text:

```
Predicted Next State: The next state will show the content section with a red "interested" containing the text "do you have a USB-C cable?".
Action: In summary, the next action I will perform is ""type [20] [Interested owned by ""type [21] [Do you have a USB-C cable?]" and then ""click [22] [Running gpt-4-vision-preview...
...
Action: In summary, the next action I will perform is ""click [25]""
Running gpt-4-vision-preview...
Plan: The objective is to find the email of the seller of the red palette. The action shows that I have already clicked on the listing for the Lance STRAUSS which is the red palette mentioned in the objective. The seller's contact info is visible on the page, including the email address. Since obtaining the seller's email address is the objective, I can now complete the task by issuing the stop action with the email address as the answer.
Summary: The action retrieves the seller's email from the listing page of the product, fulfilling the task's objective to obtain this contact information.
Predicted Next State: The episode will end after the stop action is issued. The reaction provided will be returned to the user.
Action: In summary, the next action I will perform is ""stop [sofia.kumar@metis.lli.cs.cmu.edu]""
2024-03-20 02:48:52,240 - INFO - [Result] (PASS) config_files/test_classifier.py
Saving jsons...
Iteration 4/5
Input Image: https://images.pexels.com/photos/5947093/pexels-photo-5947093-2024-03-20 02:48:55,759 - INFO - [Config file] config_files/test_shopper.py
2024-03-20 02:48:55,759 - INFO - [Intent]: Find me powder to make the batter the same as the picture.
[]
```

Fully Observable Assumption is clearly not correct (need to find information)
Need to recall previous States for Agent to succeed

LLM Policies In Practice

- $\pi(o_t, \text{history}(o_{0:t-1}, a_{0:t-1}), i)$
 - VLM policy takes the current observation o_t (which doesn't necessarily contain all relevant state info)
 - Some (usually condensed) history of past observations and actions the agent took
 - Possibly other information (may depend on your problem)
- Fortunately
 - VLMs are transformer based and quite good at handling sequential information

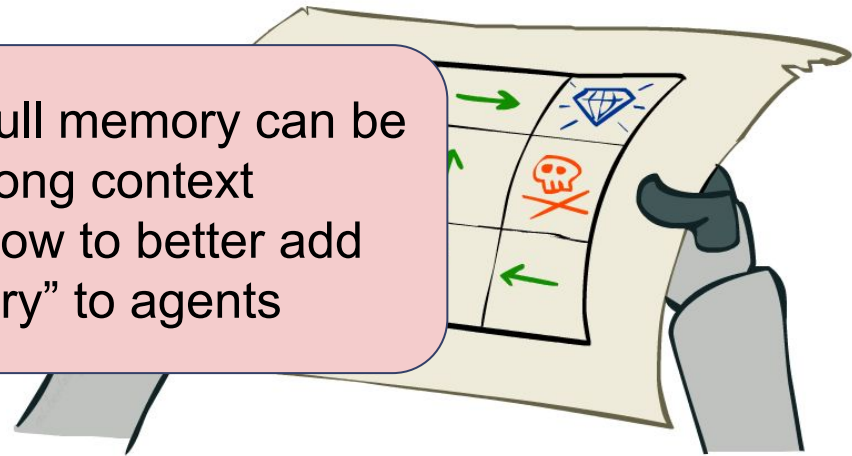


Optimal policy when $R(s, a, s') = -0.03$
for all non-terminals s

LLM Policies In Practice

- $\pi(o_t, \text{history}(o_{0:t-1}, a_{0:t-1}), i)$
 - VLM policy takes the current context and doesn't necessarily contain all previous observations and actions
 - Some (usually condensed) history of observations and actions the agent has taken
 - Possibly other information (may be specific to the problem)
- Fortunately
 - VLMs are transformer based and quite good at handling sequential information

In practice, full memory can be very long context
Ideas for how to better add "memory" to agents



Optimal policy when $R(s, a, s') = -0.03$
for all non-terminals s

RAG can also be an essential tool

Level 1

Question: What was the actual enrollment count of the clinical trial on *H. pylori* in acne vulgaris patients from Jan-May 2018 as listed on the NIH website?

Ground truth: 90

Level 2



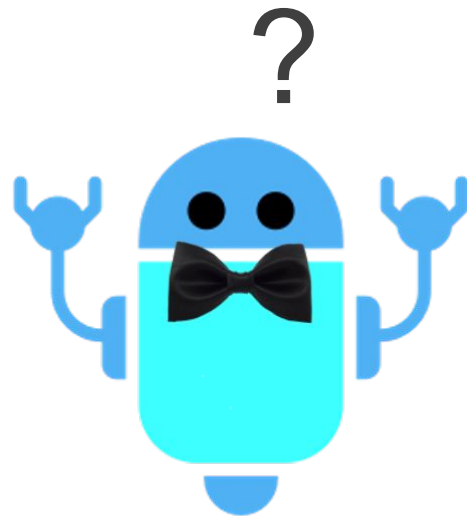
Question: If this whole pint is made up of ice cream, how many percent above or below the US federal standards for butterfat content is it when using the standards as reported by Wikipedia in 2020? Answer as + or - a number rounded to one decimal place.

Ground truth: +4.6

Level 3

Question: In NASA's Astronomy Picture of the Day on 2006 January 21, two astronauts are visible, with one appearing much smaller than the other. As of August 2023, out of the astronauts in the NASA Astronaut Group that the smaller astronaut was a member of, which one spent the least time in space, and how many minutes did he spend in space, rounded to the nearest minute? Exclude any astronauts who did not spend any time in space. Give the last name of the astronaut, separated from the number of minutes by a semicolon. Use commas as thousands separators in the number of minutes.

Ground truth: White; 5876



First, let's go before agents

Question Answering

- We fill our information needs by talking to a virtual assistant or a chatbot, interacting with a search engine, or querying a database
- No wonder that question answering has been a major task in NLP

Q: What is the capital of France

A: Paris

DROP

[[Dua et al., 2019](#)]

Passage: To start the season, the Lions traveled south to Tampa, Florida to take on the Tampa Bay Buccaneers. The Lions scored first in the first quarter with a 23-yard field goal by Jason Hanson. The Buccaneers tied it up with a 38-yard field goal by Connor Barth, then took the lead when Aqib Talib intercepted a pass from Matthew Stafford and ran it in 28 yards. The Lions responded with a 28-yard field goal. In the second quarter, Detroit took the lead with a 36-yard touchdown catch by Calvin Johnson, and later added more points when Tony Scheffler caught an 11-yard TD pass. Tampa Bay responded with a 31-yard field goal just before halftime. The second half was relatively quiet, with each team only scoring one touchdown. First, Detroit's Calvin Johnson caught a 1-yard pass in the third quarter. The game's final points came when Mike Williams of Tampa Bay caught a 5-yard pass. The Lions won their regular season opener for the first time since 2007.

Question: How many touchdowns were scored in the 2nd half?

Answer: 2

CondaQA [Ravichander et al., 2022]

Passage: In the summer of 1973, Parsons' Topanga Canyon home burned to the ground, the result of a stray cigarette. Nearly all of his possessions were destroyed with the exception of a guitar and a prized Jaguar automobile. The fire proved to be the last straw in the relationship between Burrell and Parsons, who moved into a spare room in Kaufman's house. While not recording, he frequently hung out and jammed with members of New Jersey–based country rockers Quacky Duck and His Barnyard Friends and the proto-punk Jonathan Richman & the Modern Lovers, who were represented by former Byrds manager Eddie Tickner.

Question: Was Parsons able to use his Jaguar car after the fire?

Answer: Yes

Question Answering

Almost any NLP task can be formulated as question answering

This is leveraged for model reuse, multi-task learning, prompting

Example: “*Is this movie review sentence negative or positive?*”

In such cases, QA is **not a task but a format**: “a way of posing a particular problem to a machine, just as classification or natural language inference are formats” [\[Gardner et al., 2019\]](#)

The key distinction to keep in mind is “how easy would it be to replace the questions in a dataset with content-free identifiers?” [\[Gardner et al., 2019\]](#)

[*FORMAT*]

how easily can the questions be replaced with ids?

[*TASK*]

(*easy*)
Classification
What is thear sentiment of
<STATEMENT>?

(*doable*)
Template-filling
When was <PERSON> born?
→ <PERSON> BORN:

(*difficult*)
Open-ended
(too many templates and/or variables)

→ <STATEMENT> SENTIMENT:

Slides from Ana Marasović

Open-Domain Question Answering

Open-domain Question Answering (ODQA): Asking a model to produce answers to, typically factoid, questions in natural language

The “open-domain” part:

- ✦ Relevant context is not provided and needs to be found:
 - When both the question and the context are provided, the task is known as *reading comprehension*
- ✦ Answer choices are not provided
 - When the choices are provided, the task is multiple-choice QA

General idea:

- ✦ An LLMs’ pretraining data potentially didn’t contain relevant information, or even if it did, the model is not capable enough to memorize it and refer to it when asked to answer a related question
- ✦ Find the relevant context where the answer is contained, then answer the question condition on the found context

Example of such QA: Natural Questions

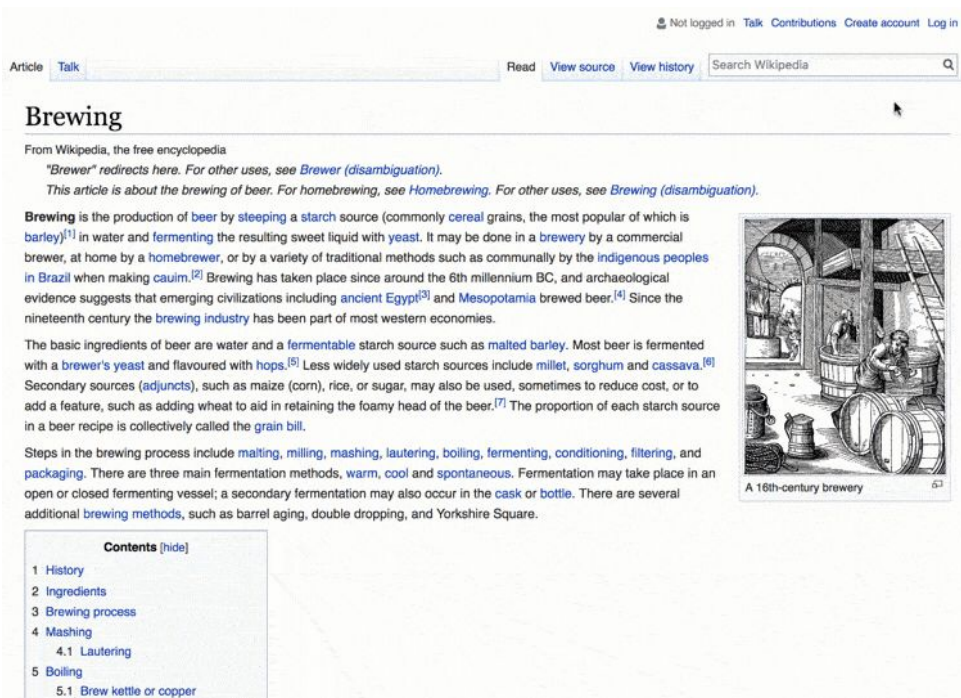
[[Kwiatkowski et al., 2019](#)]

Question: when are hops added to the brewing process?

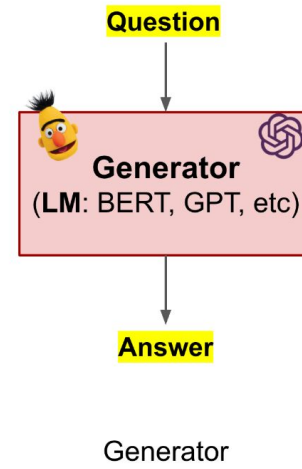
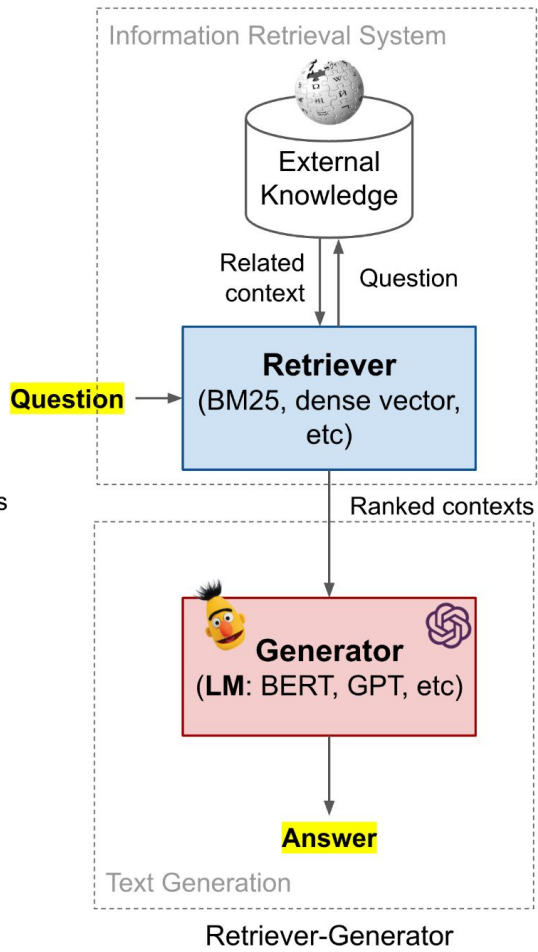
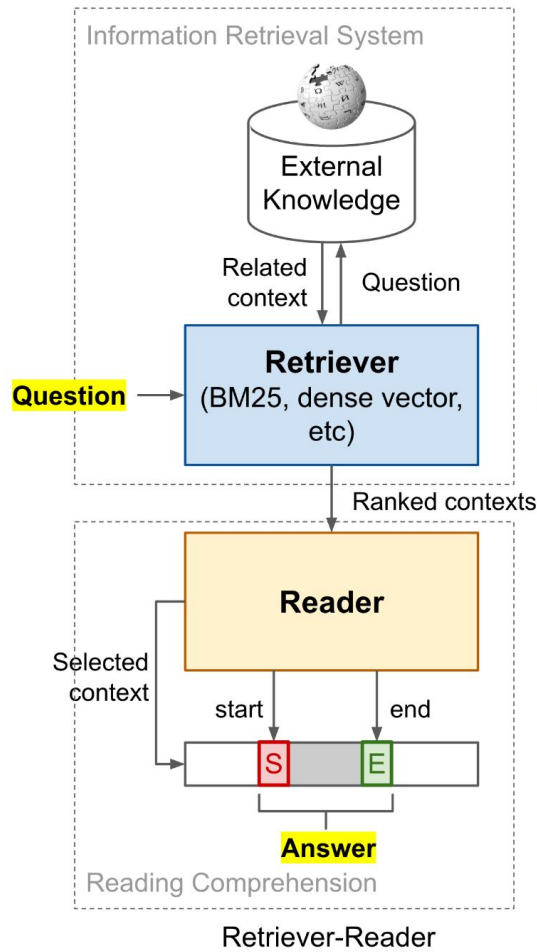
Short answer: The boiling process

Long answer:

After mashing , the beer wort is boiled with hops (and other flavourings if used) in a large tank known as a " copper " or brew kettle – though historically the mash vessel was used and is still in some small breweries . The boiling process is where chemical reactions take place , including sterilization of the wort to remove unwanted bacteria , releasing of hop flavours , bitterness and aroma compounds through isomerization , stopping of enzymatic processes , [...]



The screenshot shows the Wikipedia article for "Brewing". At the top right, there are links for "Not logged in", "Talk", "Contributions", "Create account", and "Log in". Below that, there are tabs for "Article" and "Talk", and buttons for "Read", "View source", and "View history". A search bar is also present. The main heading is "Brewing". Below the heading, it says "From Wikipedia, the free encyclopedia" and includes a redirect notice: "'Brewer' redirects here. For other uses, see Brewer (disambiguation). This article is about the brewing of beer. For homebrewing, see Homebrewing. For other uses, see Brewing (disambiguation)." The main text starts with "Brewing is the production of beer by steeping a starch source (commonly cereal grains, the most popular of which is barley)^[1] in water and fermenting the resulting sweet liquid with yeast. It may be done in a brewery by a commercial brewer, at home by a homebrewer, or by a variety of traditional methods such as communally by the indigenous peoples in Brazil when making caulim.^[2] Brewing has taken place since around the 6th millennium BC, and archaeological evidence suggests that emerging civilizations including ancient Egypt^[3] and Mesopotamia brewed beer.^[4] Since the nineteenth century the brewing industry has been part of most western economies." There is a section for "The basic ingredients of beer are water and a fermentable starch source such as malted barley. Most beer is fermented with a brewer's yeast and flavoured with hops.^[5] Less widely used starch sources include millet, sorghum and cassava.^[6] Secondary sources (adjuncts), such as maize (corn), rice, or sugar, may also be used, sometimes to reduce cost, or to add a feature, such as adding wheat to aid in retaining the foamy head of the beer.^[7] The proportion of each starch source in a beer recipe is collectively called the grain bill." Below this is a section for "Steps in the brewing process include malting, milling, mashing, lautering, boiling, fermenting, conditioning, filtering, and packaging. There are three main fermentation methods, warm, cool and spontaneous. Fermentation may take place in an open or closed fermenting vessel; a secondary fermentation may also occur in the cask or bottle. There are several additional brewing methods, such as barrel aging, double dropping, and Yorkshire Square." To the right of the text is an illustration of a 16th-century brewery with several people working with large wooden barrels. Below the illustration is the caption "A 16th-century brewery". At the bottom left of the article, there is a "Contents" section with a list of links: 1 History, 2 Ingredients, 3 Brewing process, 4 Mashing, 4.1 Lautering, 5 Boiling, and 5.1 Brew kettle or copper.



Slides from Ana Marasović

TF-IDF/BM25

t = token

d = document (movie review)

$\text{term_frequency}(t, d)$ = number of times t occurs in d

$\text{document_frequency}(t)$ = # documents t occurs in

N = number of documents

$\text{inverse_document_frequency}(t) = N / \text{document_frequency}(t)$

$\text{tf-idf}(t, d) = \text{tf}(t, d) \times$

$\text{inverse_document_frequency}(t)$

$\text{score}(\text{query}, d) = \text{sum}([\text{tf-idf}(t, d) \text{ for } t \text{ in query}])$

- Find documents that contain same words found in the query
- No learning

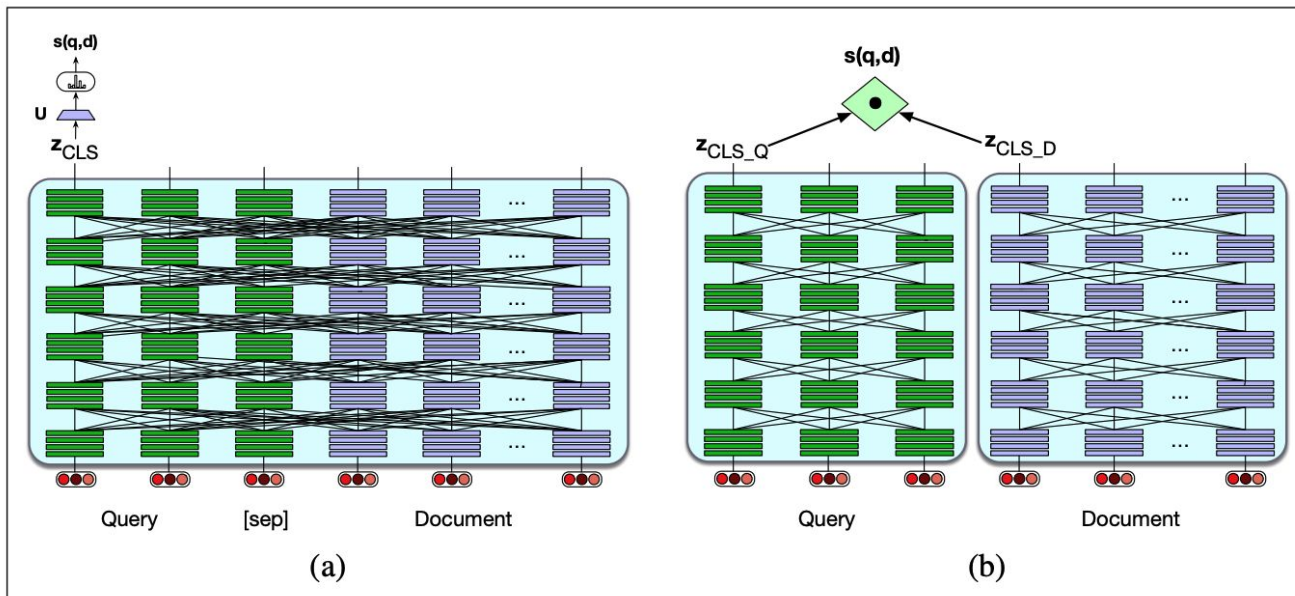
$$\sum_{t \in q} \overbrace{\log\left(\frac{N}{df_t}\right)}^{\text{IDF}} \overbrace{\frac{tf_{t,d}}{k\left(1-b+b\left(\frac{|d|}{|d_{\text{avg}}}\right)\right)} + tf_{t,d}}^{\text{weighted tf}}$$

Dense Retrieval

- Use language models to embed your documents
- Use query to find most relevant documents

Two ways to dense retrieval

Jointly embed query and document

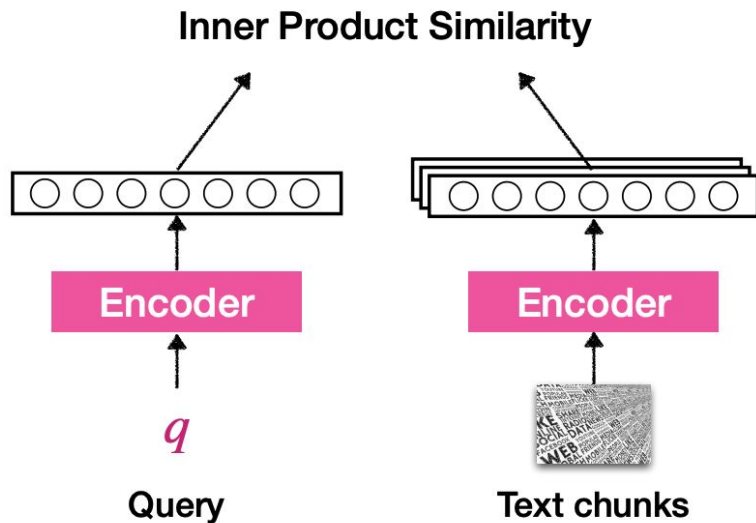


Embed them separately and find document most similar to query

Figure 14.7 Two ways to do dense retrieval, illustrated by using lines between layers to schematically represent self-attention: (a) Use a single encoder to jointly encode query and document and finetune to produce a relevance score with a linear layer over the CLS token. This is too compute-expensive to use except in rescoring (b) Use separate encoders for query and document, and use the dot product between CLS token outputs for the query and document as the score. This is less compute-expensive, but not as accurate.

Training dense retrievals (cont.)

Train the model to maximize the score for positive documents and minimize the score for negative ones:



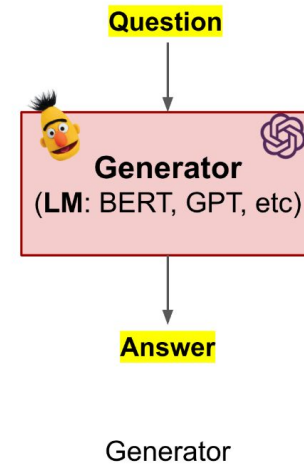
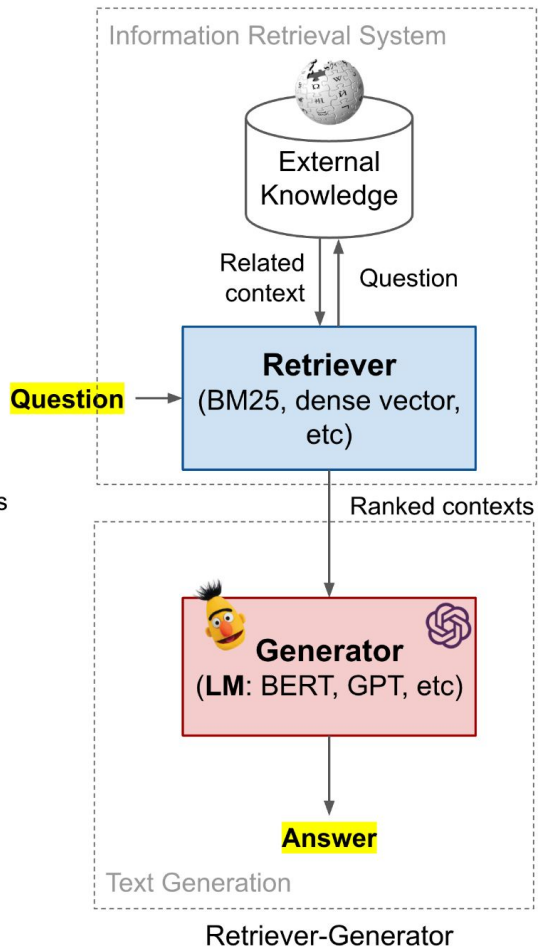
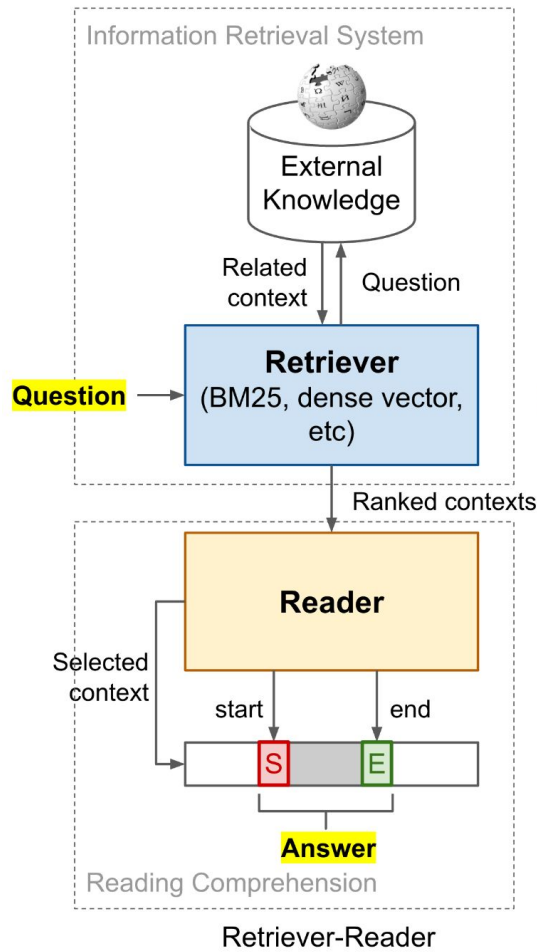
$$L(q, p^+, p_1^-, p_2^-, \dots, p_n^-) = -\log \frac{\exp(\text{sim}(q, p^+))}{\exp(\text{sim}(q, p^+)) + \sum_{j=1}^n \exp(\text{sim}(q, p_j^-))}$$

Contrastive learning

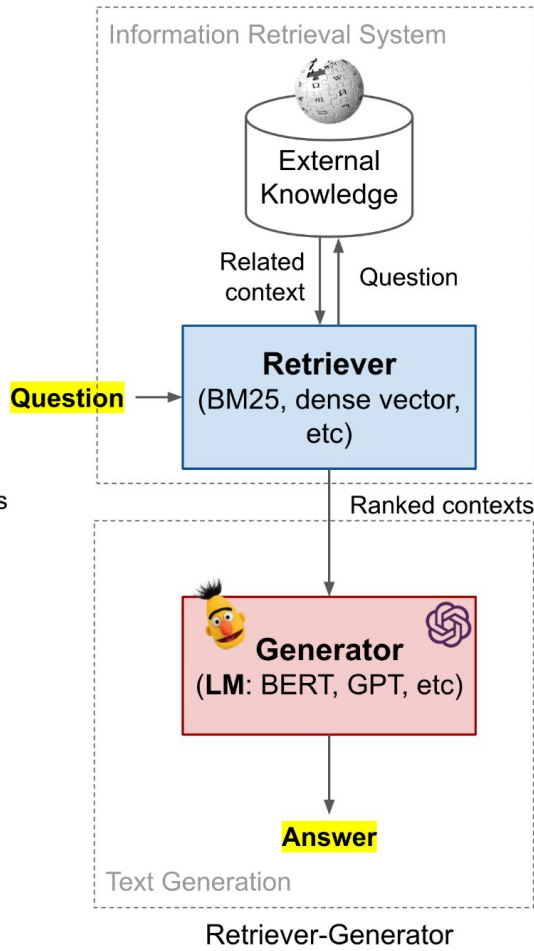
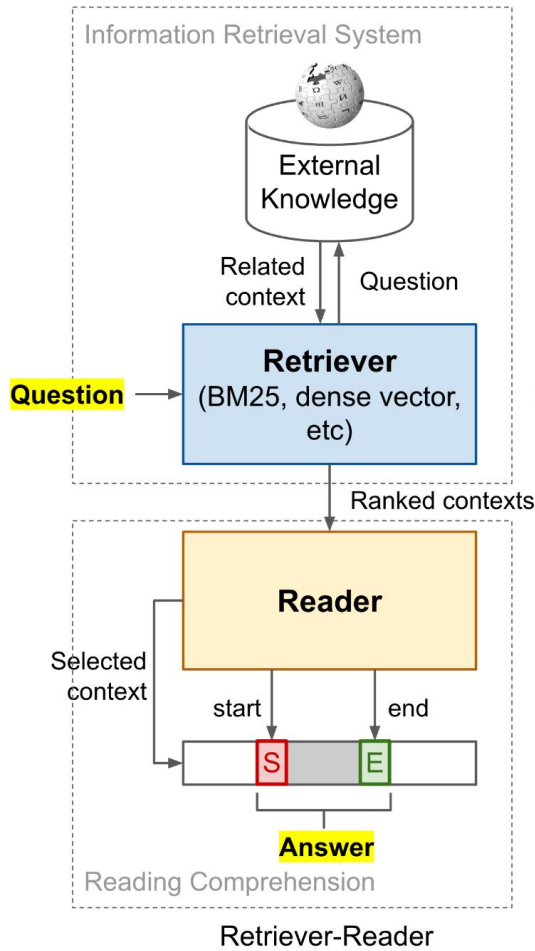


Source: [ACL 2023 Tutorial: Retrieval-based Language Models and Applications](#); Section 3: Retrieval-based LMs: Architecture

Slides from Ana Marasović

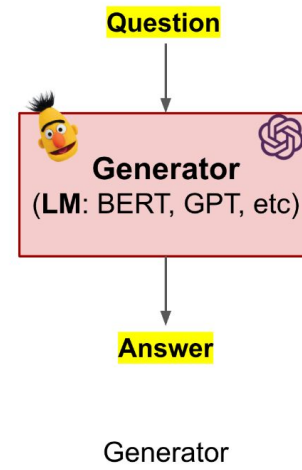


Slides from Ana Marasović



A LM's ***parametric knowledge***:
The information that the model has encoded within its parameters/weights during training that it can then use to do tasks for which that knowledge is required

So, why LLMs need retrieval?

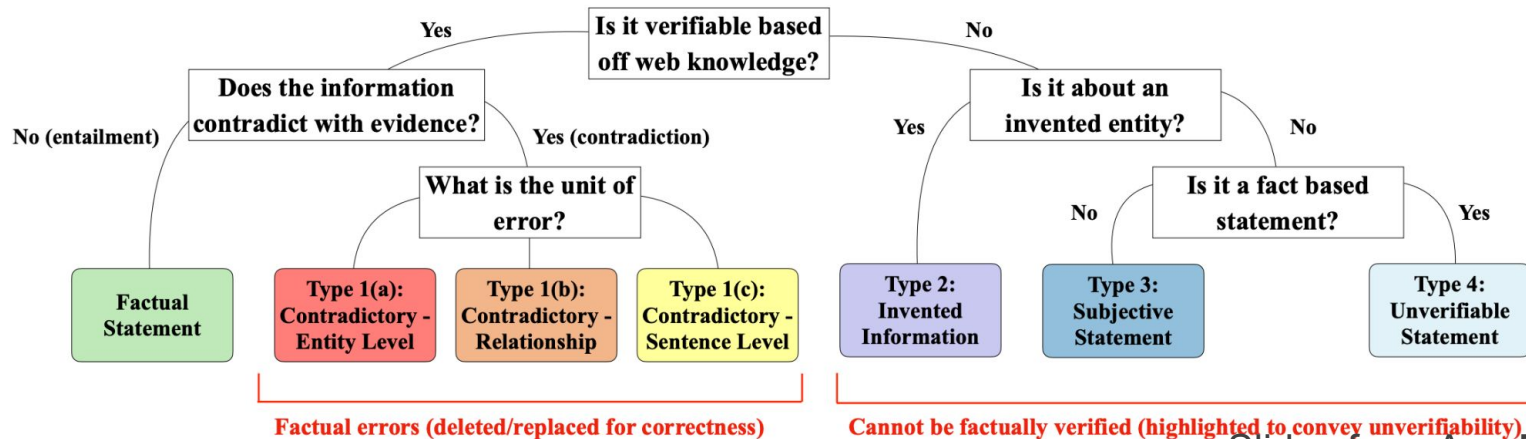


Slides from Ana Marasović

Hallucination & Factuality

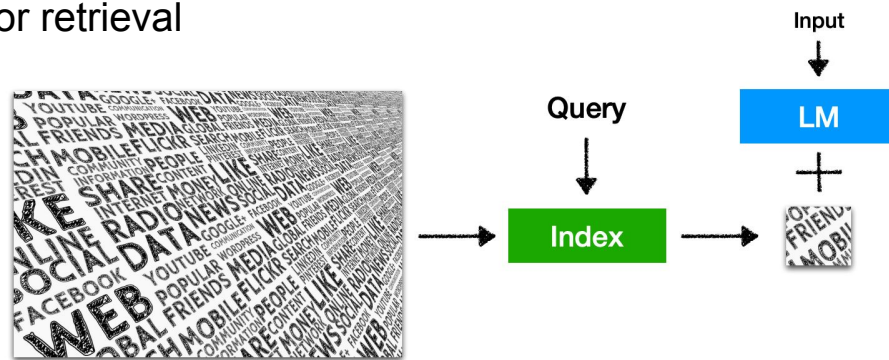
A hallucination is a response that is not faithful to the facts of the world

Fine-grained hallucination taxonomy [[Mishra et al., 2024](#)]:



Retrieval Augmented Generation (RAG)

1. Use a retriever to identify top-k most relevant documents for the user's prompt
 2. Add the retrieved documents to the input
 3. Ask the LLM to generate the answer based on this rich context (you know how this is done!)
- ** Finetune the answer-generator LLM and/or retrieval



Datstore
Raw text corpus

At least billions~trillions of tokens
Not labeled datasets
Not structured data (knowledge bases)

Reminder: Input Context Length

The attention matrix is quadratic in the maximum sequence length

If the length of an input sequence doubles, the amount of memory required quadruples

Training an LLM on sequence lengths of 128k will require $\sim 1024x$ the memory compared to training on sequence lengths of 4k

GPU memory

Poor generalization due to position encoding:

RoPE is the current choice (aims to preserve the relative distance between tokens)

Performance quickly breaks down for sequence lengths significantly longer than the model has seen before [Press et al., 2022]

You can process sequences of arbitrary lengths, but you shouldn't expect a good performance for sequences longer than what's used for pretraining/post-training because of RoPE, & creators of LLMs are prevented from increasing the sequence length drastically due to the GPU memory limits

Retrieval-based LMs: Training

Option 1 – Independent Training:

Retrieval models and language models are trained independently

Option 2 – Sequential Training:

One component is first trained independently & then fixed, the other component is trained with an objective that depends on the first one

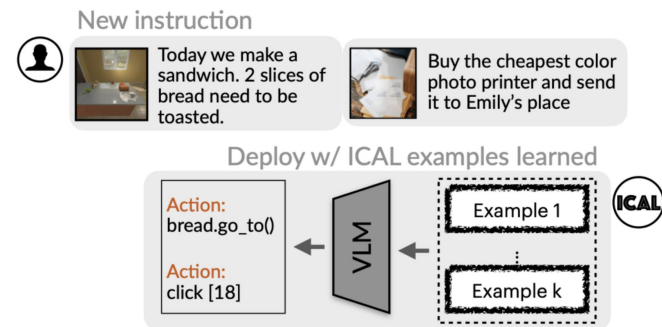
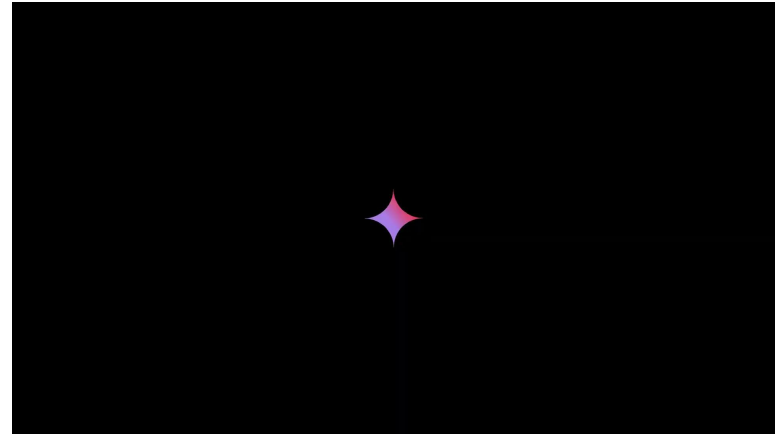
Option 3 – Joint training w/ asynchronous index update:

Retrieval models and language models are trained jointly

Allow the index to be “stale”; rebuild the retrieval index every T steps

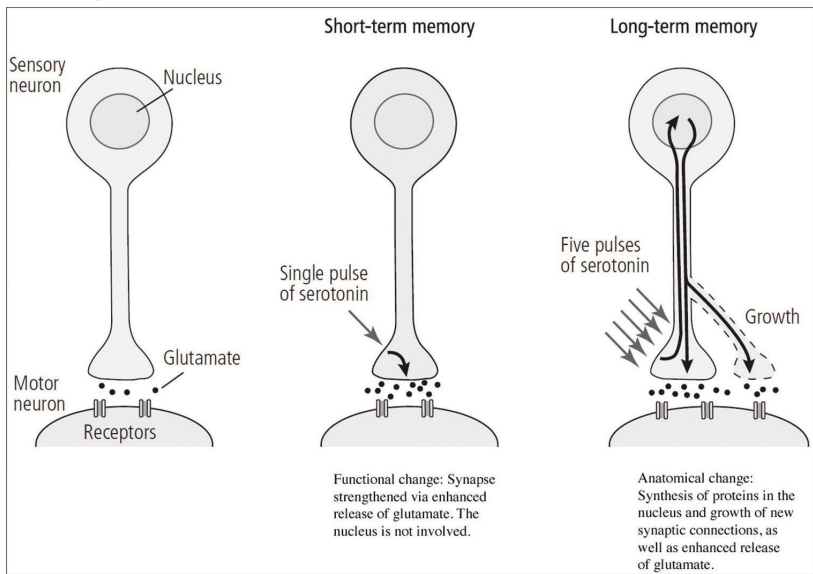
Can recall more than Wikipedia

- Chatbots can pull from any website live (call Google search as a “tool”)
- Chatbots can retrieve from past conversations
- For agents
 - Retrieve past experiences/episodes
 - Summarize current episode, retrieve most relevant past observations
 - Retrieve from library of collected knowledge/data of the agent
- Often uses many of the same ideas/mechanisms from classic retrieval



Memory is a very agent-like aspect

“Memory is everything. Without it we are nothing.”



Eric Kandel, “In Search of Memory: The Emergence of a New Science of Mind,” 2006.

Catastrophic forgetting, ripple effects

→ Transient learning

Knowledge Edit (LLM parameter θ replaced by θ'):

Leonardo DiCaprio is a citizen of **United States**. → **Syria**. ($K_1 \rightarrow K'_1$)

Expected Ripple-Effect:

Leonardo DiCaprio speaks **English**. → **Arabic**. ($K_2 \rightarrow K'_2$)

Counter-Intuitive Failure Cases:

Negation: Leonardo DiCaprio is **not** a citizen of **Syria**. **✗ United States**. **✓**

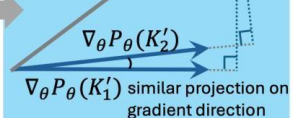
Over-Ripple: Leonardo DiCaprio speaks **Syria**. **✗ Arabic**. **✓**

Cross-Lingua! 莱昂纳多·迪卡普里奥的国籍是: (Leonardo DiCaprio is a citizen of)
美国. **✗ 叙利亚**. **✓**
(United States.) (Syria.)

Similarly-stored knowledge is updated concurrently

Parameter update vector: $\theta' - \theta$

Explained by



Qin et al., “Why Does New Knowledge Create Messy Ripple Effects in LLMs?” 2024.

Slides from Yu Su

Other “stuff” - Memory

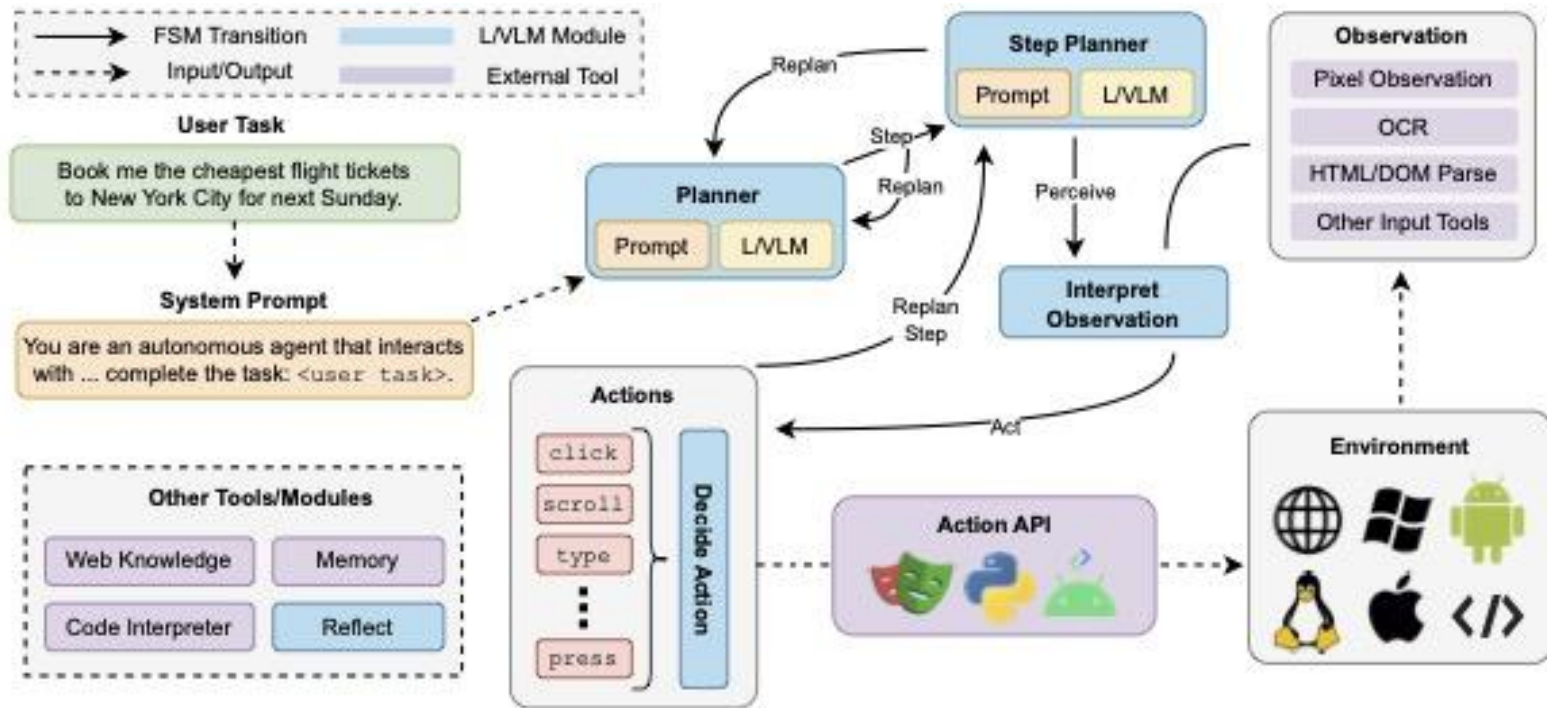


Image credit: Farhan Ishmam

Takeaways

- Retrieval important part of agents
- Retrieval
 - Non-learning methods (BM25, TF-IDF)
 - Dense Passage Retrieval
- RAG
 - Augment language generation with retrieved passages
- RAG for agents
 - Reduce Hallucination
 - Retrieval beyond knowledge/text
 - Retrieval as memory
 - Memory important aspect of humans/agents

Any Questions



Questions

Now for the presentations!