**How to hand annotate images with ReACT Reasoning Steps**

A typical example of ReACT Reasoning (with image) chain looks like this -

|  |  |  |
| --- | --- | --- |
| Image | Question | Ground Truth |
| Picture | What may be the best type of licorice? | australian |

The Usual Instructions provided to the LLM along with the image and the question at inference–

Solve a question answering task with interleaving Thought, Action, Observation. Thought can reason about the current situation, and Action can be four types:

(1) ImageObserve[image], which returns an in-depth explanation of the image, in a single paragraph.

(2) Search[entity], which searches the exact entity on Wikipedia and returns the first paragraph if it exists. If not, it will return some similar entities to search.

(3) Lookup[keyword], which returns the next sentence containing keyword in the current passage.

(4) Finish[answer], which returns the answer and finishes the task.

Please return the answer as a single entity and not a phrase or multiple outputs.

Here are some examples.

The text is followed by few-shot examples of Image/Question/Answer Triplets to guide the LLM to reason in this way.

A typical rollout in this case looks like this (hand crafted to solve the question)

Thought 1: I need to closely examine the image to identify any references to licorice, as well as potential clues that might indicate its quality.\n

Action 1: ImageObserve[image]\n

Observation 1: In the image, I notice a pack of licorice labeled 'Licorice Pieces.', but I do not yet have enough information to determine the best type of licorice.\n

Thought 2: I should consider other items in the image that might help establish the geographical context. There are receipts, books, and other items like a backpack, which could potentially give more information.\n

Action 2: ImageObserve[image]\n

Observation 2: I see one Cadbury Dairy milk triple decker chocolate, one receipt of Coles, one holy bible, a book titled “Neverwhere”. The label on another receipt and bag are non-readable due to the quality of the image.\n

Thought 3: Cadbury, Bible and the Book can be found anywhere in the world, however Coles is a supermarket chain and we can probably find out its origin and get to know which place these things (including the licorice pack) might be purchased from. \n

Action 3: Search[Coles Supermarkets]\n

Observation 3: Coles Supermarkets Australia Pty Ltd, doing business as Coles, is an Australian supermarket, retail and consumer services chain, headquartered in Melbourne as part of Coles Group. Founded in 1914 in the suburb of Collingwood by Sir George Coles, Coles currently operates 846[2] supermarkets throughout Australia, including several now re-branded Bi-Lo stores. Coles has over 120,000 employees[2][3] and accounts for around 27 per cent of the Australian market.[4]. Coles Online is the company's online shopping ('click & collect' and home delivery) service.. Between 1986 and 2006, Coles Supermarkets was a brand of Coles Myer, later Coles Group, prior to Wesfarmers purchasing Coles Group in 2007. \n

Thought 4: Oh this makes sense, most probably the photographer has bought this licorice from Coles in Australia, and since he considered this to be the best, the answer should be that australian licorice is the best.\n

Action 4: FInish[australian]

The idea is to look at the image and the question and manually, without using any AI tools, try to answer the question with Thought, Action and Observation tags. The Action consists of multiple “tools” such as

* “ImageObserve” – This is a typical example of observing an image and providing a longform caption / explanation with respect to the given question. Most of the times simpler image-question pairs can be solved directly at this step, which might not be the case in more complex datasets such as MMMU Dataset.
* “Search” – Basically it is a wrapper around going to the url – **“https://en.wikipedia.org/w/index.php?search=” .** The idea is to search an entity (search=”entity” ) on wikipedia and get its first paragraph if the entity exists. If the entity does not exist on Wikipedia simply return the top 5 suggestions provided by Wikipedia as a list. For example as shown in Figure 1, “Franklama” as an entity does not exist, and hence I should just return “Could not find Franklama. Similar: ['Frankland’, ‘HM Prison Frankland’, ‘Howard Frankland Bridge’, ‘Rosemarie Frankland’, ‘Thomas Frankland’]. In case, the entity is found on Wikipedia (for example <https://en.wikipedia.org/w/index.php?search=Coles_Supermarkets>, which exists) just return the first heading of the page as seen in Figure 2. Look at the annotated example (highlighted in green above) for the response.
* “Lookup” – A lot of times the information sought by the LLM is not present in the first paragraph of the Wikipedia article that the “Search” tool has suggested, hence a precise keyword search needs to be done. It is like going to a Wikipedia article and Ctrl+F-ing (looking up) a keyword in the article and reading that content. A lookup list is constructed by the ReACT framework if the question is not answered using the “Search” Step. A good example can be seen in the table below. Check Figure 3 for knowing how does Lookup work.
* Ultimately the answer is returned as an Action – Finish and the answer is written in the brackets – Finish[answer].

|  |  |  |
| --- | --- | --- |
| Image | Question (and answer) | React Rollout |
| Picture | What is the year this car was released? (GT Answer – 1995) | Thought 1: I need to look at the image and find out which car is this and what is it’s model year.  Action 1: ImageObserve[image]:  Observation 1: I see! The image seen here is of the car Toyota Tacoma. I cannot see the release year written anywhere though.  Thought 2: I need to determine the model year of this Toyota Tacoma. To do this, I should search for information about the Toyota Tacoma.  Action 2: Search[Toyota Tacoma]  Observation 2: The Toyota Tacoma is a pickup truck manufactured by Japanese automobile manufacturer Toyota. The Tacoma was Motor Trend's Truck of the Year for 2005. As of 2015, the Tacoma was sold in the United States, Canada, Mexico, Costa Rica,[1] Bolivia, Bermuda, and the French overseas collectivity of New Caledonia. Most markets across the world receive the Toyota Hilux in lieu of the Tacoma..  Thought 3: It does not specify a particular year of release or launch. I need to determine if the question is asking for the first model year or a specific model year mentioned in the context.  Action 3: Lookup[launch]  Observation 3: (Result 1 / 35) The Tacoma was introduced in the US in February 1995 (with a market launch in March 1995) as a replacement for the [Toyota Pickup](https://en.wikipedia.org/wiki/Toyota_Pickup) (which was the name used for the Hilux in the North American market).  Thought 3: The observation confirms that the first-generation Toyota Tacoma model was launch in 1995.  Action 4: Finish[1995] |

**A screenshot of a computer

Description automatically generated**

Figure 3 How does the Lookup function work

**A close-up of a document

Description automatically generatedA screenshot of a computer

Description automatically generatedSome things to make sure while annotating –**

Figure 2 What to do when entity is found on Wikipedia

Figure What to do when the entity is not found on Wikipedia

* I expect the annotations (after compilation) into a format shown in the link - <https://github.com/choprahetarth/mlm-benchmarking/blob/main/ReAct/annotations/aokvqa_annotated_validation.txt> . Subtleties include adding a “\n” character after every line.
* There is no need to include Search/Lookup tools for every example. However, making direct jumps using ImageObserve is not a good thing to do.
* Try to wrap up the interactions in 7 reasoning turns.
* **Here are some examples hand-annotated by me (with images)–**

[**https://github.com/choprahetarth/mlm-benchmarking/blob/main/ReAct/annotations/hand\_annotation\_aokvqa\_train.xlsx**](https://github.com/choprahetarth/mlm-benchmarking/blob/main/ReAct/annotations/hand_annotation_aokvqa_train.xlsx)

[**https://github.com/choprahetarth/mlm-benchmarking/blob/main/ReAct/annotations/hand\_annotation\_aokvqa\_validation.xlsx**](https://github.com/choprahetarth/mlm-benchmarking/blob/main/ReAct/annotations/hand_annotation_aokvqa_validation.xlsx)

[**https://github.com/choprahetarth/mlm-benchmarking/blob/main/ReAct/annotations/Prompts-FS.xlsx**](https://github.com/choprahetarth/mlm-benchmarking/blob/main/ReAct/annotations/Prompts-FS.xlsx)

[**https://uofi.app.box.com/folder/285248210280**](https://uofi.app.box.com/folder/285248210280)

**Please let me know if you are having problems in annotation**