A Implementation of AVIS workflow

We implemented AVIS using the code snippet referenced in Code 7. Throughout our experiments, we employed the APIs of Google Search, LENS, PALI, and PALM directly, without the need for additional GPU/TPU computational resources. Tools that didn’t require input queries, such as object detection, captioning, and image search, had their results pre-calculated over the two datasets to reduce the time cost. Other services like VQA, text search, and LLM QA were called during runtime.

B Comparison to pure Autonomous baseline without Transition Graph

One of the significant contributions of this paper lies in the use of a transition graph, synthesized from an authentic user study. To underscore the importance of this graph, along with user prompts in facilitating the efficacy of AVIS, we devised a baseline that operates independently of the transition graph. In this scenario, the model, at each timestep, is presented with a comprehensive list of all tools, each paired with a task description. This baseline shares similarities with the recently launched AutoGPT \(^2\), BabyAGI\(^3\) projects, which attempted to utilize LLMs as autonomous agents to select all possible actions available in the web.

The results are show in Table 3 on Infoseek WikiData unseen entity set and OKVQA. Note that this baseline doesn’t achieve the number as high as AVIS with the transition graph and user prompts. The key reason for this discrepancy is the global characteristics inherent in the tool list we have. For instance, we typically first address the visual sub-question through object detection and image search, followed by resolving the knowledge component via Google Search and LLM. However, solely relying on the task description, devoid of human behavior as guidance, can result in the model generating unrealistic tools. We will discuss this intuition more in the following sections.

<table>
<thead>
<tr>
<th>Model</th>
<th>Infoseek</th>
<th>OKVQA</th>
</tr>
</thead>
<tbody>
<tr>
<td>AVIS w.o/ Transition Graph</td>
<td>38.2</td>
<td>47.3</td>
</tr>
<tr>
<td>AVIS w/ Transition Graph</td>
<td>50.7</td>
<td>60.2</td>
</tr>
</tbody>
</table>

Table 3: Ablation of AVIS with or without the guidance of Transition Graph

C Analysis of AVIS’s generated tool execution sequence

We have also conducted an analysis to determine whether common patterns exist within the generated programs of AVIS’s predictions.

We gathered the tool execution traces for all samples within the Infoseek unseen entity dataset. Initially, we display the frequency of each tool being invoked in Figure 6, followed by a more detailed analysis of the first to fourth most commonly called tools in Figures 7-10. As illustrated, the AVIS model, guided by the transition graph and prompts, does not utilize all possible combination of tools, but favors some certain combinations. For instance, as depicted in Fig 7, "object select" is utilized more frequently than other tools at the outset. Similarly, as demonstrated in Fig 9, during the third step, when the model accumulates the visual answer, it is likely to invoke "web search" to gather additional information.

We have also calculated the transition probability of the induced graph in Fig 11. The structure of this graph differs slightly from the guided transition graph because during actual runtime, the model will not predict some of the edges. Overall, it reveals a clear two-step question-solving pattern. Initially, AVIS gathers sufficient visual information through the use of visual tools such as "object detection," "VQA," or "identical image search," and then employs "LLM QA" to obtain the visual answer. Subsequently, it iteratively calls "web search" and "LLM QA" post-search with a prompt, eventually deriving the final answer. We also present the distribution of the lengths of generated sequences in Figure 13. As illustrated, the lengths vary considerably, rather than maintaining a fixed value, with a length of 5 being most common for the generated sequences.

\(^2\)https://github.com/Significant-Gravitas/Auto-GPT
\(^3\)https://github.com/yoheinakajima/babyagi
Another intriguing aspect worth exploring is our reasoner component. As explained in the paper, the reasoner evaluates whether the output of each tool is "informative," "not informative," or "answerable". We exhibit the overall frequency of these predictions in Figure 12. As shown, the model tends to classify most of the outputs as either informative or answerable. However, approximately 8.1% of returned entries are deemed "not informative," in which case AVIS would backtrack to select alternative actions. We further demonstrate a few examples of different choices in Table 4.

D Dataset Details

Infoseek\(^4\) is a Visual Question Answering (VQA) dataset, specifically geared towards information-seeking questions that cannot be answered merely through common sense knowledge. This dataset was curated by initially gathering human-annotated questions, which were then automatically integrated with existing visual entity recognition datasets and Wikidata to generate complex question-

\(^4\)https://open-vision-language.github.io/infoseek/
Table 4: Several examples of API execution results and the reasoner’s justification.

answer pairs. At the time of submission, we only have access to its wikidata split. Here we also report the results on human split in Table 5.

OK-VQA is another VQA dataset, unique in its requirement for the application of external knowledge that transcends the information directly visible in the input images. The creation of this dataset involved crowdsourced workers who were tasked with annotating complex questions, drawing upon the extensive knowledge resources available on Wikipedia.

E Prompt Examples

Below we show different prompt examples to support our AVIS workflow. First is the prompts for planning, which selects which tool to use and what query to send. It is consists of a overall task descriptions and many real examples showing at which circumstances real users select this tool.

https://okvqa.allenai.org/
Figure 11: Induced transition frequency graph of AVIS over Infoseek dataset.

```python
task_instructions = {
    'vqa': 'You will ask simple question about this image to a external QA module. Please use this when the input query is very straightforward and simple.',
    'object_select': 'You will select one of the object we detect to dig further. Please use when the question asks about a specific object.',
    'identical_image_search':
}
```
Figure 12: Overall frequency of judgement by reasoner of AVIS.

Figure 13: Length distribution of AVIS’s generated action sequences.

<table>
<thead>
<tr>
<th>Model</th>
<th>Unseen Entity</th>
<th>Unseen Question</th>
</tr>
</thead>
<tbody>
<tr>
<td>PALM (Q-only, few-shot)</td>
<td>6.6</td>
<td>4.8</td>
</tr>
<tr>
<td>OFA (fine-tune)</td>
<td>2.9</td>
<td>6.2</td>
</tr>
<tr>
<td>PALI (fine-tune)</td>
<td>5.9</td>
<td>13.3</td>
</tr>
<tr>
<td>PALM w/ CLIP (few-shot + external knowledge)</td>
<td>14.9</td>
<td>15.6</td>
</tr>
<tr>
<td>FiD w/ CLIP (fine-tune + external knowledge)</td>
<td>17.6</td>
<td>18.9</td>
</tr>
<tr>
<td>AVIS (ours, few-shot)</td>
<td>31.4</td>
<td>33.6</td>
</tr>
</tbody>
</table>

Table 5: **Visual Question Answering** results (accuracy) on Infoseek$_{human}$. The first four rows are results from their paper that do not use external knowledge, and the next two are from their paper that use CLIP as knowledge source.

```
17
21 "You will see captions of all images identical to the given image. Please use when the question asks about the whole image instead of a part."
22 'image_search':
23 'You will see captions of all images similar to this object. Please use when you need more information.'
24 'web_search':
25 'You will send question to Google Search to get knowledge. Please use when the current query requires extra knowledge'
26 'llm_qa':
27 'You will send question to a QA module. Please use this when the input query is simple and contain common-sense knowledge'
28 }

Listing 1: Planner prompt skeleton and Task instructions
vqa_plan_prompts = [

  **Query:** what is the train carrying?
  **Context:**
  a train traveling down train tracks next to a forest. There are four trains on the railway track. In the background there are trees, poles and sky. (Caption, whole image)
  **Extracted Text:** BNSF (score=100.0),
  BNSF Railway: BNSF Railway is one of the largest freight railroads in North America (score=89.3),

  **Action:** vqa

**Query:** What is the girl wearing on her legs?
**Context:**
 a woman standing in a field putting on a coat. There is a woman standing on the ground. This is grass and there are plants. In the background we can see some trees and this is sky. (Caption, whole image)

**Query:** what color is the bus?
**Context:**
a double decker bus parked in front of a building. There is a double decker bus on the road and this is snow. Here we can see a pole, light, trees, and houses. In the background there is sky. (Caption, whole image)
  **Extracted Text:** ENVIRO400 (score=100.0),
  Extracted Text: Les Miserables (score=100.0),
  **Query Suggestion:** les miserables (score=100.0),
  Volvo Olympian: The Volvo Olympian was a rear-engined 2-axle and 3-axle double decker bus chassis manufactured by Volvo at its Irvine, Scotland factory (score=88.5),
  Alexander Dennis Enviro400: The Alexander Dennis Enviro400 is a twin-axle low-floor double-decker bus that was built by the British bus manufacturer Alexander Dennis between 2005 and 2018 (score=85.4),

  **Action:** vqa

**Query:** what is the person doing?
**Context:**
two people sitting on the floor opening presents. There are sofas on the sofas there are pillows, here there is table, on the table there are plants and other objects, here there are two persons sitting on the ground, gift boxes, dog and this is floor. (Caption, whole image)

object_select_plan_prompts = [

  **Query:** what is the name of this building?
  **Context:**
  a group of people that are standing in front of a building. There is a building in the left corner which has few people standing in front of it and there is a fire hydrant in the right corner and there is a street light pole beside it. (Caption, whole image)
  **Query Suggestion:** Alcatraz Warden's House San Francisco (score=100.0),
  Alcatraz Island (historic_site): Alcatraz Island is a small island! (score=91.9),
  Warden's House: The Warden's House was the home of the wardens of the federal penitentiary on Alcatraz Island, off San Francisco (score=78.1),

  **Action:** object_select

**Query:** what is the island?
**Context:**
a view of a mountain from a cable car. There is a ropeway. Behind that there are trees and hills. (Caption, whole image)
  **Query Suggestion:** Ngong Ping 360 (gondola_lift_station): Ngong Ping 360 is a bicable gondola lift on Lantau Island in Hong Kong (score=91.8),
  Tian Tan Buddha (monument): The Big Buddha is a large bronze statue of Buddha, completed in 1993, and located at Ngong Ping, Lantau Island, in Hong Kong (score=79.0),

  **Action:** object_select

**Query:** what is the name of this place?
**Context:**
a cemetery with a building in the background. There is a road and there are many atoms and trees beside it and there is a building in the right corner. (Caption, whole image)

**Query:** what is the name of this bird?
**Context:**
a bird sitting on top of a lush green hillside. There is a bird on the grassland in the foreground area of the image and the background is blurry. (Caption, whole image)
  **Query Suggestion:** Atlantic puffin (type_of_bird): The Atlantic puffin, also known as the common puffin, is a species of seabird in the auk family (score=73.2),
  Horned puffin (type_of_bird): The horned puffin is an auk found in the North Pacific Ocean, including the coasts of Alaska, Siberia and British Columbia (score=73.2),
  Puffins (type_of_bird): Puffins are any of three species of small alcids in the bird genus Fratercula (score=73.2),
  Fraterculinae (score=48.8),
  Auk (type_of_bird): An auk or alcid is a bird of the family Alcidae in the order Charadriiformes (score=48.8),

  **Action:** object_select

**Query:** what is the name of this building?

identical_image_search_plan_prompts = [

**Query:** what is the name of this building?
**Context:**
 a double decker bus parked in front of a building. There is a double decker bus on the road and this is snow. Here we can see a pole, light, trees, and houses. In the background there is sky. (Caption, whole image)

  **Query Suggestion:** Alcatraz Warden's House San Francisco (score=100.0),
  Alcatraz Island (historic_site): Alcatraz Island is a small island! (score=91.9),
  Warden's House: The Warden's House was the home of the wardens of the federal penitentiary on Alcatraz Island, off San Francisco (score=78.1),

  **Action:** object_select

**Query:** what is the island?
**Context:**
a view of a mountain from a cable car. There is a ropeway. Behind that there are trees and hills. (Caption, whole image)
  **Query Suggestion:** Ngong Ping 360 (gondola_lift_station): Ngong Ping 360 is a bicable gondola lift on Lantau Island in Hong Kong (score=91.8),
  Tian Tan Buddha (monument): The Big Buddha is a large bronze statue of Buddha, completed in 1993, and located at Ngong Ping, Lantau Island, in Hong Kong (score=79.0),

  **Action:** object_select

**Query:** what is the name of this place?
**Context:**
a cemetery with a building in the background. There is a road and there are many atoms and trees beside it and there is a building in the right corner. (Caption, whole image)

**Query:** what is the name of this bird?
**Context:**
a bird sitting on top of a lush green hillside. There is a bird on the grassland in the foreground area of the image and the background is blurry. (Caption, whole image)
  **Query Suggestion:** Atlantic puffin (type_of_bird): The Atlantic puffin, also known as the common puffin, is a species of seabird in the auk family (score=73.2),
  Horned puffin (type_of_bird): The horned puffin is an auk found in the North Pacific Ocean, including the coasts of Alaska, Siberia and British Columbia (score=73.2),
  Puffins (type_of_bird): Puffins are any of three species of small alcids in the bird genus Fratercula (score=73.2),
  Fraterculinae (score=48.8),
  Auk (type_of_bird): An auk or alcid is a bird of the family Alcidae in the order Charadriiformes (score=48.8),

  **Action:** object_select

**Query:** what is the name of this building?

identical_image_search_plan_prompts = [

**Query:** what is the name of this building?

identical_image_search_plan_prompts = [

**Query:** what is the name of this building?

identical_image_search_plan_prompts = [

**Query:** what is the name of this building?

identical_image_search_plan_prompts = [

**Query:** what is the name of this building?
A group of people that are standing in front of a building. There is a building in the left corner which has few people standing in front of it and there is a fire hydrant in the right corner and there is a street light pole beside it. (Caption, whole image)

Query Suggestion: Alcatraz Warden's House San Francisco (score=100.0),
Alcatraz Island (historic_site): Alcatraz Island is a small island 1 (score=91.9),
Warden's House: The Warden's House was the home of the wardens of the federal penitentiary on Alcatraz Island, off San Francisco (score=78.1),

Action: identical_image_search

Query: what is the aircraft?
Context: A fighter jet sitting on top of an airport tarmac. There is a plane and missiles on the ground. At the left a person is standing wearing a cap. (Caption, whole image)
Extracted Text: AIRLINERS.NET (score=100.0),
Query Suggestion: airliners.net (score=100.0),
Airliners: Airliners (score=74.8),
British Aerospace Hawk 200: The British Aerospace Hawk 200 is a single-seat, single engine light multirole fighter designed for air defence, air denial, anti-shipping, interdiction, close air support, and ground attack (score=74.8),
product: Airfix BAE Hawk T1 1:72 (score=0.0),
product: Rolls-royce Adour In The Hawk / Bae Hawk 200 . Pdf/download (score=0.0),

Action: identical_image_search

Query: what is the name of this place?
Context: A row of pillars sitting next to a dirt road. There is a building and this is plant. Here we can see pillars and a sky. (Caption, whole image)
Query Suggestion: Palmyra Archaeology (score=100.0),
Great Colonnade at Palmyra (ancient_roman_architecture_structure): The Great Colonnade at Palmyra was the main colonnaded avenue in the ancient city of Palmyra in the Syrian Desert (score=90.3),

Action: identical_image_search

Query: what is the name of this lake?
Context: A view of a river surrounded by mountains. There are trees in the right corner and there is a river and mountains in front of it. (Caption, whole image)
Monte Bre (peak): Monte Bre is a small mountain east of Lugano on the flank of Monte Boglia with a view of the bay of Lugano and the Pennine Alps and the Bernese Alps (score=85.5),
product: Top Searched (score=0.0),

Action: identical_image_search

We then show how AVIS decompose question into a visual sub-question and a knowledge sub-question. This is done at beginning to guide later tool usage.

```python
question_decomposition_prompt = """"Read the following question for a given image. Decompose the question into two sub-questions.
The first will ask information about the image, and the second requires reasoning over the textual knowledge.
In the second question, we use # to denote the answer of the first question.

Question: What chemical makes the vegetable orange?
Visual: which orange vegetable is shown?
Knowledge: chemical makes # orange?

Question: How long can their horns grow?
Visual: which animals are shown?
Knowledge: How long can #’s horns grow?

Question: What is a competition for these animals called?
Visual: which animals are shown?
Knowledge: competition for #?

Question: What is the name of the ancient greek sport that evolved into the sport featured above?
Visual: which sport is played?
Knowledge: name of the ancient greek sport that evolved into #?

Question: Which food item here has the most protein?
Visual: what are the food items shown?
Knowledge: Which food item of # has the most protein?

Question: How many calories are in this meal?
```
```
Object select prompt = """"

Below are several examples to help AVIS learns how to select the most suitable object ID.

1. Query: which city is this place?
   Object #0 [a row of pillars sitting next to a dirt road. There is a building and this is plant. Here we can see pillars and a sky. (Caption, whole image)]
   Query Suggestion: Palmyra Archaeology (score=100.0),
   Great Colonnade at Palmyra (ancient_roman_architecture_structure): The Great Colonnade at Palmyra was the main colonnaded avenue in the ancient city of Palmyra in the Syrian Desert (score=90.3),
   Object #1 [a green plant sitting next to a brick wall. There is a plant and this is wall. And there is a sky. (Caption, center)]
   Date palm (type_of_palm_trees): Phoenix dactylifera, commonly known as date palm, is a flowering plant species in the palm family, Arecaceae, cultivated for its edible sweet fruit called dates (score=81.7),
   European rabbit (type_of_leporids): The European rabbit or coney is a species of rabbit native to the Iberian Peninsula, western France, and the northern Atlas Mountains in northwest Africa (score=31.3),
   GreenView Fairway Formula Seed Success Paillis biodegradable avec engrais Sac de 4,5 kg Couvre 200 m2 (score=0.0),
   Willow: Willows, also called sallows and osiers, of the genus Salix, comprise around 350 species of typically deciduous trees and shrubs, found primarily on moist soils in cold and temperate regions (score=31.3),
   Tamarisk: The genus Tamarix is composed of about 50-60 species of flowering plants in the family Tamaricaceae, native to drier areas of Eurasia and Africa (score=26.8),
   Output: The query asks about the city of the place. Only Object #0 (whole image) mentions city name Palmyra, which is an ancient city. Also, Object #0 contains Query Suggestion "Palmyra Archaeology".
   Therefore, the predicted Object #ID is 0.

2. Query: where is this place?
   Object #0 [a view of a valley surrounded by mountains. There are hills and this is grass. Here we can see trees and this is sky. (Caption, whole Image)]
   Monterey Pine (type_of_conifers): Pinus radiata, the Monterey pine, insignis pine or radiata pine, is a species of pine native to the Central Coast of California and Mexico (score=49.1),
   European rabbit (type_of_leporids): The European rabbit or coney is a species of rabbit native to the Iberian Peninsula, western France, and the northern Atlas Mountains in northwest Africa (score=31.3),
   Willow: Willows, also called sallows and osiers, of the genus Salix, comprise around 350 species of typically deciduous trees and shrubs, found primarily on moist soils in cold and temperate regions (score=31.3),
   Tamarisk: The genus Tamarix is composed of about 50-60 species of flowering plants in the family Tamaricaceae, native to drier areas of Eurasia and Africa (score=26.8),
   GreenView Fairway Formula Seed Success Paillis biodegradable avec engrais Sac de 4,5 kg Couvre 200 m2 (score=0.0),
   Output: The query asks about the location of this place. Although these entries don't explicitly contain location name, but Object #1 (center) contains Monterey Pine and European rabbit, which might hint the location later.
   Therefore, the predicted Object #ID is 1.

3. Query: What type of sandwich is this?
   Visual: a sandwich with meat and cheese.
   Knowledge: #

4. Question: What is the main ingredient in this food?
   Visual: which food is shown?
   Knowledge: main ingredient in #?
Below are the prompts to extract answer from objects and extracted captions of similar images.

reason_vqa_prompt = """"Please think step by step. In the following, you will be given:
1. Query: The query to be asked.
2. Think: Why the following knowledge is retrieved.
3. Entity: A list of entities that describe the object.
4. Retrieval: A list of web documents that are similar to the object. If there’s "high confidence", it’s very important.
5. Your task is to predict a short answer to the query based on the provided information. You need to first identify which knowledge entry is mostly relevant, and then extract the answer from the knowledge.
6. Help: If the object contains "query suggestion", try to use it. Otherwise, if a information appears lots of time, there’s a higher chance it’s the answer.
7. After explaining your decision choice, saying "Answer is" and appending your predicted short answer. Please also generate the type of the answer after a comma.

Query: what is this building?
Think: object (whole image) contains stockholm city hall, which is the seat of stockholm municipality in stockholm, sweden.
Object: [ Stockholm City Hall (city_hall): Stockholm City Hall is the seat of Stockholm Municipality in Stockholm, Sweden (score=96.1). 
Bla Hallen (banquet_hall): The Blue Hall is the main hall of the Stockholm City Hall best known as the banquet hall for the annual Nobel Banquet, and also used for state visits, student balls, jubilees and other large events (score=79.0).
]
Retrievals: [ high confidence: City Hall - Blue Hall (1) | Stockholm (2) | Pictures | Sweden in Global-Geography (score=47.8), high confidence: le salon bleu a city hall (salle de remise des prix nobel) - Picture of Stockholm, Stockholm County - Tripadvisor (score=47.7), ]
Output: The query asks about the building. From both Object and Retrievals, there are mentions about Stockholm City Hall and Blue Hall. As Stockholm City Hall contains Blue Hall, the answer shall be Stockholm City Hall.
Therefore, the predicted answer is Stockholm City Hall.

Query: which sport is played?
Think: object shows a snail sitting on top of a tennis ball.
Object: [ Cantareus apertus (type_of_gastropods): Cantareus apertus, commonly known as the green garden snail, is a species of air-breathing land snail, a terrestrial pulmonate gastropod mollusc in the family Helicidae, the typical snails, Garden snail (type_of_gastropods): Cornu aspersum, known by the common name garden snail, is a species of land snail in the family Helicidae, which includes some of the most familiar land snails, Slug: Slug, or land slug, is a common name for any apparently shell-less terrestrial gastropod mollusc, Snail: A snail is a shelled gastropod, ]
]
Output: The query asks about sport. From both entities and retrievals, they only talk about snail, and there is no information about which sport is played.
Therefore, given the provided information, this query cannot be answered.
Output: The query asks about which sport is played. From retrievals, there exist many mentions about netball, and mentions that they are played by women. Therefore, the predicted answer is women netball.

Query: What is the name of the insect?
Think: Only object (while image) mentions the name of the insect, Western Tiger Swallowtail.
Object: {
Query Suggestion: Western Tiger Swallowtail (score=100.0),
Canadian Tiger Swallowtail (type_of_lepidoptera): Papilio canadensis, the Canadian tiger swallowtail, is a species of butterfly in the family Papilionidae (score=78.4),
Eastern Tiger Swallowtail (us_state_butterfly): Papilio glaucus, the eastern tiger swallowtail, is a species of butterfly native to eastern North America (score=78.4),
}
Retrievals: {
high confidence: Kupu-kupu – Wiktionary (score=100.0),
high confidence: Top Spots for Nature Watching and Birding | VisitMaryland.org (score=100.0),
high confidence: File:Eastern Tiger Swallowtail Papilio glaucus on Milkweed 2800px.jpg - Wikimedia Commons (score=99.8),
high confidence: Photographing Butterflies - Life in the Finger Lakes (score=97.8),
}
Output: The query asks about the name of the insect. From Object, it contains a very informative "Query Suggestion: Western Tiger Swallowtail". Therefore, the predicted answer is Western Tiger Swallowtail.

Listing 5: Reason Prompt (Visual Question)

Below are prompts AVIS extract answer from search results:

reason_qa_prompt = ***
Please think step by step. In the following, you will be given a "Query", and a list of "Knowledge" from Google Search related to this query.

Your task is to predict a short answer to the query based on the provided information. You need to first identify the most relevant knowledge entry, and then predict a short answer based on the knowledge. If a information appears lots of time, there's a higher chance it's the answer.

After explaining your decision choice, saying "Answer is" and appending your predicted answer.
If you are uncertain about the answer, especially when the knowledge is irrelevant to the query, say "cannot be answered". Do not generate the answer not inside the provided knowledge.

Query: What chemical makes carrot orange?
Knowledge: {
Title: How did carrots become orange? - The Economist
Content: High Confidence Response: carotenoids.
Context: The chemical compounds that give carrots their vivid colour, carotenoids, are usually used by plants that grow above ground to assist in the process of photosynthesis.

Title: Why are carrots orange? | Ask Dr. Universe | Washington State University
Content: Orange carrots are packed with chemicals called carotenoids—specifically, beta-carotene. Your body turns beta-carotene into vitamin A, which helps you grow and protects you from getting sick.

Beta-carotene isn’t just nutritious. It’s also loaded with orange pigment. That’s why vegetables with lots of beta-carotene—like sweet potatoes, squash, and pumpkins—share the same color. But what about that rainbow of other carrot colors? They have their own special qualities, too. Purple carrots get their color from
}
Output: The query asks about chemical that makes carrot orange. Because there's one high confidence result, the most relevant knowledge entries about such chemical is "High Confidence Response: carotenoids."
From this result we know the chemical shall be carotenone.
Therefore, the predicted answer is carotenone.

Query: What is the name of the drainage basin of Ounasjoki?
Knowledge: {
Title: Ounasjoki – Wikipedia
}

Query: What is the name of the drainage basin of Ounasjoki?
Knowledge: {
Title: Ounasjoki – Wikipedia
}

Query: What is the name of the drainage basin of Ounasjoki?
Knowledge: {
Title: Ounasjoki – Wikipedia
}
Therefore, the predicted answer is Martina Bergman-Osterberg.

Course. The Ounasjoki originates at Ounasjarvi lake in Enontekiö. It flows first eastwards through Peräjärvi lake and转弯 south after some seven kilometres. The river then follows southern-su

Title: DRAINAGE BASIN OF THE BALTIC SEA - UNECE

Content: It is also the largest river entirely within its borders. Ounasjoki is approximately 295.6 kilometre (186.2 mi) in length, and the catchment area is 23,958 square kilometres (5,393 sq mi), 27% of the Kemijoki catchment area.

Tributaries

- Nakkalajoki.
- Kakkolajoki.
- Syvä Tapastajoki.
- Loukinen.
- Meltauksjoki.

Course. The Ounasjoki originates at Ounasjarvi lake in Enontekiö. It flows first eastwards through Peräjärvi lake and turns south after some seven kilometres. The river then follows southern-su

Title: DRAINAGE BASIN OF THE BALTIC SEA - UNECE

Content: It is also the largest river entirely within its borders. Ounasjoki is approximately 295.6 kilometre (186.2 mi) in length, and the catchment area is 23,958 square kilometres (5,393 sq mi), 27% of the Kemijoki catchment area.

Tributaries

- Nakkalajoki.
- Kakkolajoki.
- Syvä Tapastajoki.
- Loukinen.
- Meltauksjoki.

...
The query asks about number of elevators in Torre Picasso. The most relevant knowledge entry about number of elevators in Torre Picasso is "26 elevators; 18 serve office floors divided into three zones:"

From the result, we know the number of elevators shall be 26.

Therefore, the predicted answer is 26.
F Error Analysis

Figure 14: Error Type: LLM planning module miss important details. Question: What country does this building belong to? Correct answer: Austria; Predicted answer: United States

- Step 1: object_select, query = What is this building?
- Step 2: image_search, query = what is this building?
  Retrievals: [ Millennium tower wien hi-res stock photography and images - Alamy (score=48.4), Millennium Tower Office Building in Vienna, Austria.....]
  Reason: The query asks about the building. From Object, there is a mention about "Millennium Tower" with high score. From Retrievals, there are many mentions about "Millennium Tower" with high score. Therefore, the predicted answer is Millennium Tower, building.
- Step3: Google Search, query = country of Millennium Tower, building?
  Title: Number 8: Millennium Tower in SoMa, San Francisco - San Francisco YIMBY
  Content: The eighth tallest skyscraper in the Bay Area planned or built is the Millennium Tower...
  From the search result, "The Millennium Tower is a skyscraper in San Francisco, California", the predicted answer is United States.
  yes, answerable, Output = United States

- image_search query=what is this bridge?
  Retrievals: high confidence: Na Drini çuprija - Zadužbina Mehmed-paše Sokolovića - Furaj.ba | S nama u avanturu (score=87.2), high confidence: File:NKD279 Most Mehmed Paše Sokolovića Višegrad 5.jpg - Wikimedia Commons (score=91.3), ......]
Figure 15: **Error Type**: LLM QA module extracts wrong evidence. **Question**: What is the architectural style of this bridge? **Correct answer**: Ottoman architecture; **Predicted answer**: turkish monumental architecture

yes, answerable. The query asks about the bridge. From Retrievals, there are many mentions about "Mehmed Paša Sokolović Bridge". Therefore, the predicted answer is Mehmed Paša Sokolović Bridge, bridge.

- **Google Search**, query=architectural style of Mehmed Paša Sokolović Bridge, bridge?
  Title: Mehmed Pasha Sokolovic Bridge – Višegrad, Bosnia and Herzegovina - Atlas Obscura
  Content: Celebrated example of Ottoman architecture that also served as a site for brutal massacres. Been Here? ......
  yes, answerable. Question asks about architectural style of Mehmed Paša Sokolović Bridge, bridge. From the search results, the most relevant knowledge entry about architectural style is "It is characteristic of the apogee of Turkish monumental architecture and civil engineering". Therefore, the predicted answer is Turkish monumental architecture.

- **Step 1**: identical_image_search, query = What is this place? Retrievals: [ high confidence: Reportage entre les murs de Mariama Bâ, internat d’excellence pour jeunes filles sénégalaises - Debout Congolaises (score=95.3), ] not informative, backtrack
- **Step 2**: object_select, query = What is this place? the query asks about the place. the image contains boats, water, trees and buildings. there is no explicit information about the place. therefore, the predicted object id is 0, which is whole image.
  Step 3: image_search, query = What is this place? Object: [ a group of boats floating on top of a body of water. There are two white boats on the water. Behind that there are trees and buildings. (Caption, whole image) ] Retrievals: [ high confidence: Reportage entre les murs de Mariama Bâ, internat d’excellence pour jeunes filles sénégalaises - Debout Congolaises (score=95.3), ] not informative, backtrack
- **Step 4**: vqa, query = What is this place? output=a lake yes, answerable
- **Step 5**: llm_qa, query = body of water is a lake located in or next to? output=land
Figure 16: Error Type: Tool provides incorrect information. Question: Which body of water is this place located in or next to? Correct answer: Atlantic; Predicted answer: land