

---

## Appendix

---

### A Supplementary Material

#### A.1 Pseudo-code of Workforce

Define the following symbols:  $\mathcal{F}$  denotes the failure count,  $\mathcal{I}$  represents the failure information set,  $\mathcal{B}$  indicates the task failure status,  $\mathcal{P}$  represents the planner agent,  $\mathcal{C}$  represents the coordinator agent, and  $\mathcal{T}$  represents the trajectory set. The overall processing flow of WORKFORCE is shown in Algorithm 2.

---

##### Algorithm 1 Workforce Framework

---

**Require:** Task  $T$ , Worker Registry  $\mathcal{W}$ , Task Channel  $\mathcal{C}$ , Max Replanning Tries  $K$

**Ensure:** Final Output  $O$

```
1: Initialize:
2:  $\mathcal{F} \leftarrow 0$ 
3:  $\mathcal{I} \leftarrow \{\}$ 
4:  $\mathcal{B} \leftarrow \text{False}$ 
5: while  $\mathcal{F} \leq K$  do
6:   Planning
7:    $S \leftarrow \{\}$  ▷ /* Initialize subtask set */
8:   if  $|\mathcal{I}| > 0$  then
9:      $S \leftarrow \mathcal{P}.\text{replan}(T, \mathcal{W}, \mathcal{I})$ 
10:  else
11:     $S \leftarrow \mathcal{P}.\text{decompose}(T, \mathcal{W})$ 
12:  end if
13:  Coordinating and Processing
14:   $R \leftarrow \{\}$  ▷ /* Initialize result set */
15:  for each subtask  $s_i \in S$  do
16:     $w_i \leftarrow \mathcal{C}.\text{find\_assignee}(s_i, \mathcal{W})$  ▷ /* Worker assignment */
17:     $\mathcal{C}.\text{post}(\mathcal{C}, s_i, w_i)$  ▷ /* Post task to channel */
18:     $r_i \leftarrow w_i.\text{process\_task}(s_i)$  ▷ /* Process subtask */
19:    if  $r_i$  is Failed then
20:       $\mathcal{C}.\text{post}(\mathcal{C}, \text{Failure}(s_i))$  ▷ /* Report failure */
21:       $\mathcal{B} \leftarrow \text{True}$ 
22:       $\mathcal{I} \leftarrow \mathcal{I} \cup \{s_i.\text{failure\_reason}\}$ 
23:      break
24:    end if
25:     $R \leftarrow R \cup \{r_i\}$  ▷ /* Collect result */
26:     $\mathcal{C}.\text{post}(\mathcal{C}, r_i)$  ▷ /* Post result to channel */
27:  end for
28:  if not  $\mathcal{B}$  then
29:    break
30:  else
31:     $\mathcal{F} \leftarrow \mathcal{F} + 1$ 
32:  end if
33: end while
34:  $O \leftarrow \mathcal{P}.\text{synthesize}(R)$  ▷ /* Synthesize final output */
35: return  $O$ 
```

---

## A.2 Error Analysis Examples

### Incorrect Plan: Error Case Level3 #12

**Question:** I was referencing each of the tables in the file from papers that were cited by the “Trans fatty acid contents in chocolates and chocolate wafers in Turkey” paper. I lost my own reference sheet and need to know which of the papers each table came from. The file may not use the full table caption. If the references in the “Trans fatty acid” paper bibliography were numbered starting with 1, give me the numbers in the order that they would be used to fill the cells in the Excel file from top to bottom, as a comma separated list.

**Answer:** 8, 29, 22, 1, 8, 26

**Annotation steps:** 1. Searched “Trans fatty acid contents in chocolates and chocolate wafers in Turkey” on Google. 2.Opened [https://www.researchgate.net/publication/234034780\\_Trans\\_fatty\\_acid\\_contents\\_in\\_chocolates\\_and\\_chocolate\\_wafers\\_in\\_Turkey](https://www.researchgate.net/publication/234034780_Trans_fatty_acid_contents_in_chocolates_and_chocolate_wafers_in_Turkey). 3.Opened the Excel file. 4.Searched each reference in the paper on Google. 5.Checked any free-to-access reference for a table similar to the titles in the Excel file. 6.Added the numbers of the references to the Excel file. 7.Copied the numbers into a comma-separated list.

### Incorrect Plan Error Analysis

**Analysis:** The subtask only lets the agent match the table content with the reference paper, without further extracting the content of the reference paper itself.

**Agent Content:**

```
{“subtask”: “Write and execute code to analyze the Excel file content and match the tables with the references from the paper.”,  
“result”: “Table-to-Reference Matching AnalysisBased on my analysis of the Excel file and the bibliography from the “Trans fatty acid contents in chocolates and chocolate wafers in Turkey” paper, I’ve matched each table to its corresponding reference number.  
}
```

### Subtask Ambiguity: Error Case Level3 #18

**Question:** The year is 2022. I am at the National Air and Space Museum east of the Potomac River. I want to go to Fire Station 301 DCA ARFF using the metro. I go in the wrong direction and end up at the station closest to Cleveland Elementary School. How many metro stations am I away from my original destination if I don't change lines? Your answer should be a numerical integer value.

**Answer:** 8

**Annotation steps:** 1. Google search "National Air and Space Museum". 2. Note there are two National Air and Space Museums. One in Virginia, the other in Washington D.C. 3. Google map search "Potomac River" and zoom out. 4. See that Washington DC is east of the Potomac River. 5. Determine that the National Air and Space Museum refers to the one in Washington D.C. 6. Google search "Metro Station National Air and Space Museum Washington D.C.". 7. Clicked on the first result: Getting Here — National Air and Space Museum, <https://airandspace.si.edu/visit/museum-dc/directions>. 8. Read on the website, "The closest Metrorail stop is at L'Enfant Plaza." Note this location. 6. Google map search "Fire Station 301 DCA ARFF". 7. Zoom out to look for nearby metro stations. 8. The closest station is Ronald Reagan Washington National Airport. 9. Google map search "Cleveland Elementary School". 10. The closest metro station to Cleveland Elementary School is Shaw-Howard Univ Station. 11. Google search "DC Metro Station Map". 12. Clicked on the second result: 2022 System Map, <https://www.wmata.com/schedules/maps/upload/2022-System-Map.pdf>. 13. Locate L'Enfant Plaza station. It is the transfer station for all color lines. 14. Locate Shaw-Howard Univ stations 4 stops above L'Enfant Plaza station. 15. Locate Ronald Reagan National Airport station on the blue/yellow line. 16. Recall the current location: Shaw-Howard Univ station's yellow/green line. 17. Since the question says no line changes, we deduce the line must be one that Shaw-Howard Univ and Ronald Reagan National Airport stations have in common: yellow line. 18. Begin at Shaw-Howard Univ station and follow the yellow line. 19. Count the number of stops until it reaches Ronald Reagan National Airport station. 20. Final answer: 8.

### Subtask Ambiguity Error Analysis

**Analysis:** The subtask asks how many subway stations there are between the stations, but the question actually requires calculating how many subway stations you need to pass to reach your destination, which is different.

**Agent Content:**

```
{ "subtask": "Calculate how many metro stations are between the station nearest to Cleveland Elementary School and the station nearest to Fire Station 301 DCA ARFF, without changing lines. Provide the final answer as a numerical integer value.",  
  "result": "Calculating Metro Stations Between Shaw-Howard University and National Airport Based on the information provided in the dependency results and my analysis of the Washington DC Metro system, I need to calculate how many metro stations are between the station nearest to Cleveland Elementary School and the station nearest to Fire Station 301 DCA ARFF, without changing lines. Identifying the Stations From the dependency results:- The station nearest to Cleveland Elementary School is **Shaw-Howard University** on the Green Line The station nearest to Fire Station 301 DCA ARFF is **Ronald Reagan Washington National Airport** (also called 'National Airport') on the Blue and Yellow Lines Determining the Connecting Line According to the dependency results, the Yellow Line connects both stations. This means a person can travel between these two stations without changing lines by using the Yellow Line. Calculating the Number of Stations I analyzed the Yellow Line stations from north to south and found:- Shaw-Howard University is station 9 on the Yellow Line- National Airport is station 17 on the Yellow Line The stations between them (excluding the starting station) are: 1. Mount Vernon Square 2. Gallery Place 3. Archives 4. L'Enfant Plaza 5. Pentagon 6. Pentagon City 7. Crystal City Therefore, there are **7** metro stations between Shaw-Howard University and National Airport on the Yellow Line without changing lines. Final Answer The number of metro stations between the station nearest to Cleveland Elementary School and the station nearest to Fire Station 301 DCA ARFF, without changing lines, is **7**.",  
}
```

### Tool Selection Error: Error Case Level3 #8

**Question:** I was trying to remember how well the Cheater Beater performed in comparison to the Cheater when James tested it on his channel. I know that the Cheater still outperformed the Cheater Beater in terms of CFM. Could you please look that up for me, and report the CFM of both the Cheater and the Cheater Beater? I'm not sure if he made any changes to his testing, but this was back in season 4, so just report the value from that season. Please format your response like this: CFM number for Cheater, CFM number for Cheater beater.

**Answer:** 101.376, 84.348

**Annotation steps:** 1. Using a web browser, navigate to a search engine and conduct a search: "James Cheater Cheater Beater CFM Season 4" 2. Finding no relevant result, navigate to a search engine and conduct another search: "Cheater Beater Season 4" 3. Navigate to the first search result, <https://www.youtube.com/watch?v=2vq3COPZbKo> 4. Evaluate the YouTube page, noting that the video description identifies the video content comparing the performance of computer fans to a fan referred to as the "cheater" 5. Follow the link to the YouTube channel Major Hardware, <https://www.youtube.com/@MajorHardware> 6. Navigate to the About tab link, <https://www.youtube.com/@MajorHardware/about> 7. Evaluate the content, noting that the page identifies the operator of the channel as James 8. Navigate to a search engine and conduct a search, "James Major Hardware Cheater Beater" 9. Navigate to the first result, identical to the result from step 3 above, <https://www.youtube.com/watch?v=2vq3COPZbKo> 10. Search the page for CFM, finding no result 11. Load the video content and review it 12. Note an onscreen text element identifying a fan as "CALL SIGN: CHEATER BEATER" at timestamp 224 13. Note an onscreen table identifying the performance of various fans tested during season four, at timestamp 485 14. Evaluate the table content, identifying an entry for a fan named "Cheater" and a fan named "Cheater Beater" 15. Evaluate the table content, identifying that the data for both fans were recorded in season 4, S4E1 for Cheater, S4E6 for Cheater Beater 16. Record the data from the CFM column for the two fans, "Cheater: 101.376", and "Cheater Beater: 84.348" 17. Report the correct response to my user: "Cheater: 101.376 Cheater Beater: 84.348".

### Tool Selection Error Error Analysis

**Analysis:** The agent traverses season4's videos, but only stops at searching and does not open the video page for searching and processing.

**Agent Content:**

```
{ "role": "assistant", "tool_calls": [ { "type": "function",  
"function": { "name": "web_search",  
"arguments": { "question": "Major Hardware Fan Showdown Season 4 Episode 6 Cheater  
Beater" } } } ], },  
{ "role": "assistant", "tool_calls": [ { "type": "function",  
"function": { "name": "web_search",  
"arguments": { "question": "Major Hardware Fan Showdown Season 4 Episode 7 Cheater  
Beater" } } } ], },  
{ "role": "assistant", "tool_calls": [ { "type": "function",  
"function": { "name": "web_search",  
"arguments": { "question": "Major Hardware Fan Showdown Season 4 Episode 8 Cheater  
Beater" } } } ] }
```

### Web Toolkit Failure: Error Case Level1 #38

**Question:** On Cornell Law School website's legal information institute, under the fifth section of federal rules alphabetically, what word was deleted in the last amendment to the first rule in the article that has "witnesses" in the most titles as of 2021?

**Answer:** inference

**Annotation steps:** 1. Searched "Cornell Law School legal information institute" on Google. 2. Opened <https://www.law.cornell.edu/>. 3. Clicked Get The Law - Federal Rules - Federal Rules of Evidence (fourth section down). 4. Found the article that has "witnesses" in the most titles (VII). 5. Opened the first rule (701). 6. Scrolled to the last amendment as of 2021 (2011 amendment). 7. Found the word that was deleted (inference).

### Web Toolkit Failure Error Analysis

**Analysis:** The question asked to count the titles containing the exact word "witnesses". As a result, the browser also counted the titles containing the singular "witness" (such as "Who May Impeach a Witness").

**Agent Content:**

```
{ "role": "assistant",  
"tool_calls": [ { "type": "function",  
"function": { "name": "browse_url",  
"arguments": { "start_url": "https://www.law.cornell.edu/rules/fre/article_vi", "task_prompt": "List  
all the rule titles in Article VI of the Federal Rules of Evidence to count how many contain the  
word 'witnesses'" } } } ], },  
{ "role": "tool",  
"content": "The task was to list all the rule titles in Article VI of the Federal Rules of Evidence  
and count how many contain the word \"witnesses.\" Here's the final answer:The following rule  
titles contain the word \"witnesses\": 1. Rule 605. Judge's Competency as a Witness 2. Rule  
606. Juror's Competency as a Witness 3. Rule 607. Who May Impeach a Witness 4. Rule  
608. A Witness's Character for Truthfulness or Untruthfulness 5. Rule 611. Mode and Order of  
Examining Witnesses and Presenting Evidence 6. Rule 612. Writing Used to Refresh a Witness  
7. Rule 613. Witness's Prior Statement 8. Rule 614. Court's Calling or Examining a Witness 9.  
Rule 615. Excluding Witnesses. Total count: 9 rule titles contain the word \"witnesses.\" ", }
```

### Document Toolkit Failure: Error Case Level2 #44

**Question:** The work referenced in footnote 397 of Federico Lauria's 2014 dissertation is also the source for the titles of two paintings in the Smithsonian American Art Museum's collection, as of August 2023. What is the absolute difference between the chapter numbers of the chapters that the titles of these two paintings quote?

**Answer:** 8

**Annotation steps:** 1. Use search engine to search for "Federico Lauria's 2014 dissertation". 2. Open the result from philarchive.org and open the PDF file for the full paper. 3. Search for footnote 397 to find that the referenced work is Thomas Hobbes's "Leviathan". 4. Use search engine to search for "Smithsonian American Art Museum collection search". 5. Go to the museum's search webpage. 6. Enter "Hobbes Leviathan" into the search box and submit the search. 7. Open the two results, one by Jan Stussy ("A free man...") and one by Leon Karp ("Hereby it is manifest..."). 8. Verify from the full titles of these works that the titles are quotes from "Leviathan". 9. Use search engine to search for "Thomas Hobbes Leviathan full text". 10. Open any result that contains the full text, like the Project Gutenberg version. 11. Search the text for the titles of each painting, using different substrings from the titles as needed to account for variations in spelling and punctuation. 12. Find that the "A free man..." quote is from Chapter XXI (21) and that the "Hereby it is manifest..." quote is from Chapter XIII (13). 13. Calculate the absolute difference of the chapter numbers:  $21 - 13 = 8$ .

### Document Toolkit Failure Error Analysis

**Analysis:** The PDF address of the paper has been determined, but the Page limit exceeded.

**Agent Content:**

```
{ "role": "assistant", "tool_calls": [ { "function": { "name": "extract_document_content",  
"arguments": "{ \"document_path\": \"https://philarchive.org/go.pl?id=LAUQLO&proxyId=&u=https%3A%2F%2Fphilpapers.org%2Farchive%2FLAUQLO.pdf\", \"query\": \"footnote 397 Moby Dick chapter titles\" }" } ] },  
{ "role": "tool",  
"content": "(True, 'Error while processing document: Page limit exceeded')" }
```

### Multimodal Toolkit Failure: Error Case Level3 #17

**Question:** Which of the fruits shown in the 2008 painting "Embroidery from Uzbekistan" were served as part of the October 1949 breakfast menu for the ocean liner that was later used as a floating prop for the film "The Last Voyage"? Give the items as a comma-separated list, ordering them in clockwise order based on their arrangement in the painting starting from the 12 o'clock position. Use the plural form of each fruit.

**Answer:** pears, bananas

**Annotation steps:** 1. Use search engine to search for "2008 painting Embroidery from Uzbekistan". 2. Open the top result, a link to the painting's page on the Dayton Art Institute website, and verify that the painting has the specified title and year. 3. Identify the fruits in the painting as watermelon, pear, lemon, and banana, which can be verified by either watching the video on the page or reading its linked transcript. 4. Use search engine to search for "ocean liner floating prop The Last Voyage". 5. Note from the results that this ocean liner was the SS Ōcele de France. 6. Use search engine to search for "October 1949 breakfast menuSS Ōcele de France". 7. Go to the result that shows the vintage SS Ōcel de France breakfast menu for October 1949. 8. Search the menu for each of the four fruits from the painting, finding "Pear" and "Bananas" but no matches for "lemon" or "watermelon". 9. Check the positions of the fruits in the painting to find that the pears come before the bananas in clockwise order starting from the 12 o'clock position. 10. Format the final answer as specified using the correct ordering: pears, bananas.

### Multimodal Toolkit Failure Error Analysis

**Analysis:** ImageAnalysisToolkit recognizes that the banana is placed at 12 o'clock and the pear is placed at 9 o'clock, but their actual positions are 11 o'clock and 7 o'clock respectively.

**Agent Content:**

{ "subtask": "Create the final answer by listing the matching fruits in a comma-separated list, ordered clockwise from the 12 o'clock position as they appear in the painting, using the plural form of each fruit.",

"result": "I'll solve this task by creating a final answer that lists the matching fruits in a comma-separated list, ordered clockwise from the 12 o'clock position as they appear in the painting, using the plural form of each fruit. Based on the dependency results, I need to: 1. Identify which fruits from the painting "Embroidery from Uzbekistan" were also served on the SS le de France's October 1949 breakfast menu 2. List these matching fruits in clockwise order starting from the 12 o'clock position 3. Use the plural form of each fruit From the comparison analysis in the dependency results, I can see that the fruits that appear in both the painting and the breakfast menu are: - Bananas - Pears Now, I need to arrange these in clockwise order starting from the 12 o'clock position as they appear in the painting. According to the dependency results, the clockwise order of fruits in the painting starting from 12 o'clock is: **1. Bananas (12 o'clock) 2. Watermelon (3 o'clock) 3. Pears (6 o'clock) 4. Lemons (9 o'clock)** Since only bananas and pears match with the breakfast menu, and they are already in plural form, the final answer is: bananas, pears This is the comma-separated list of matching fruits in clockwise order from the 12 o'clock position as they appear in the painting, using the plural form of each fruit." }

### Response Format: Error Case Level2 #85

**Question:** At the two-minute mark in the YouTube video uploaded by the channel GameGrumps on May 14, 2017 as part of their playthrough of the game Mario Kart 8 Deluxe, the show hosts are competing on one of the games racetracks. What was the world record time for that track in the games 150cc mode as of June 7, 2023? Express your answer in minutes and seconds, rounding the seconds to the nearest hundredth, e.g. 1:01.001.

**Answer:** 1:41.614

**Annotation steps:** 1. Search the web for "gamegrumps mario kart 8 deluxe may 14 2017". 2. Click on the YouTube video result. 3. Navigate to two minutes into the video. 4. Scroll further back until I see the name of the racecourse, Yoshi Circuit. 5. Search the web for "mario kart 8 deluxe yoshi circuit world record 150cc". 6. Scroll down until I find a reliable world record listing site. 7. Navigate through the site until I find the record that meets the specified criteria. 8. Read the date the record was set to confirm that it applies to the question's specified date.

### Response Format Error Analysis

**Analysis:** This question requires rounding the seconds to the nearest hundred, such as 1:01.001, but the agent finally gives the result of \*\*1:41.61\*\*.

**Agent Content:**

{ "role": "assistant",

"content": "I need to analyze the results from all subtasks to determine the final answer to the original question. Looking at the results: 1. The first subtask identified the correct GameGrumps video from May 14, 2017: "Mario Kart 8 Deluxe: The Grand Prix - PART 7" (<https://www.youtube.com/watch?v=nvaLkvUkW0w>). 2. The second subtask determined that at the two-minute mark of this video, the hosts are racing on \*\*Yoshi Circuit\*\*. 3. The third subtask found that the world record time for Yoshi Circuit in 150cc mode as of June 7, 2023 was \*\*1:23.964\*\*. 4. The fourth subtask formatted this time as \*\*1:23.96\*\* (rounding to the nearest hundredth). 5. However, the verification task found a different time: \*\*1:41.614\*\* from an archived version of the Mario Kart 8 Deluxe World Records website from May 29, 2023. There's a discrepancy between the two reported world record times. The verification task provides more specific details, including: - The record holder: "Army" from France Date set: March 10, 2023



- Character/kart setup used: Birdo (Black) with Mr. Scooty kart, Azure Roller tires, and Paper Glider The verification task also specifically mentions “GCN Yoshi Circuit,” which is the full name of the track (GCN indicating it originated in GameCube’s Mario Kart: Double Dash). Given that the verification task provides more detailed information and cites a specific archive date (May 29, 2023) that is closer to our target date (June 7, 2023), I will consider this the more reliable result. Therefore, **the final answer to the original question is: 1:41.61**(1 minute, 41.61 seconds) This is the world record time for Yoshi Circuit in Mario Kart 8 Deluxe’s 150cc mode as of June 7, 2023, expressed in minutes and seconds with the seconds rounded to the nearest hundredth.” }

### Reasoning Error: Error Case Level3 #25

**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.

**Answer:** White; 5876

**Annotation steps:** 1. Use search engine to search for “NASA’s Astronomy Picture of the Day 2006 January 21”. 2. Open the link to the image. 3. Read the explanation to find that the image is of astronaut Charles “Pete” Conrad reflected in the helmet of astronaut Alan Bean. 4. Observe that the smaller astronaut in the image is the one reflected in the other’s helmet, so the smaller astronaut must be Charles “Pete” Conrad. 5. Go to the Wikipedia page for Charles “Pete” Conrad. 6. Search for “Astronaut Group” to find that Conrad was a member of NASA Astronaut Group 2. 7. Open the Wikipedia pages for each member of NASA Astronaut Group 2. 8. For those who are not deceased, go to View history and select the latest version of their Wikipedia page as of August 2023. 9. Compare the times listed in the infobox of each astronaut’s Wikipedia page under “Time in space”, observing that Ed White has the least time in space with 4d 01h 56m, but also that Elliott See does not have a listed “Time in space”. 10. Read through Elliott See’s Wikipedia article to find that he died in an accident before his first space flight, so he should be excluded, making Ed White’s 4d 01h 56m the least amount of time in space. 11. Convert 4d 01h 56m to minutes:  $4d * 24h/d * 60m/h + 1h * 60m/h + 56m = 5,876m$ . 12. Format the final answer as specified: White; 5,876.

### Reasoning Error Error Analysis

**Analysis:** The accumulated time in space for NASA’s “New Nine” had been correctly captured, but the people who had not been to space were incorrectly taken into account.

#### Agent Content:

{ “subtask”: “Research the space time duration for each astronaut in this group. Use web search to find reliable sources about each astronaut’s missions and total time spent in space. Create a comprehensive list with their names and space time durations.”,

“result”: “Space Time Durations for NASA Astronaut Group 2 Members”

1. **\*\*Neil Armstrong\*\*** Total time in space: 8 days, 14 hours, 12 minutes Missions: Gemini 8 (1966), Apollo 11 (1969)
2. **\*\*Frank Borman\*\*** Total time in space: 19 days, 21 hours, 35 minutes Missions: Gemini 7 (1965), Apollo 8 (1968)
3. **\*\*Charles Pete Conrad\*\*** Total time in space: 49 days, 3 hours, 38 minutes Missions: Gemini 5 (1965), Gemini 11 (1966), Apollo 12 (1969), Skylab 2 (1973)
4. **\*\*Jim Lovell\*\*** Total time in space: 29 days, 19 hours, 5 minutes (715 hours, 5 minutes) Missions: Gemini 7 (1965), Gemini 12 (1966), Apollo 8 (1968), Apollo 13 (1970)
5. **\*\*James McDivitt\*\*** Total time in space: 14 days, 2 hours, 56 minutes Missions: Gemini 4 (1965), Apollo 9 (1969)
6. **\*\*Elliot See\*\*** Total time in space: 0 minutes **Note: Selected but never flew; died before Gemini 9 mission.**



7. **Thomas P. Stafford** Total time in space: 21 days, 3 hours, 43 minutes (507 hours, 43 minutes) Missions: Gemini 6A (1965), Gemini 9A (1966), Apollo 10 (1969), Apollo-Soyuz Test Project (1975)
8. **Ed White** Total time in space: 4 days, 1 hour, 56 minutes Missions: Gemini 4 (1965); first American spacewalk; died in Apollo 1 fire.
9. **John Young** Total time in space: 34 days, 19 hours, 39 minutes Missions: Gemini 3 (1965), Gemini 10 (1966), Apollo 10 (1969), Apollo 16 (1972), STS-1 (1981), STS-9 (1983) }

### Language/Question Ambiguity: Error Case Level2 #73

**Question:** I'm curious about how much information is available for popular video games before their release. Find the Wikipedia page for the 2019 game that won the British Academy Games Awards. How many revisions did that page have before the month listed as the game's release date on that Wikipedia page (as of the most recent entry from 2022)?

**Answer:** 60

**Annotation steps:** 1. Search the web for British Academy Video Games Award for Best Game 2019. 2. Find the answer, Outer Wilds. 3. Find the Wikipedia page for Outer Wilds. 4. Go to the last revision from 2022. 5. Note the release date, May 29, 2019. 6. View the page history. 7. Count how many edits were made to the page before May 2019. 8. Arrive at the answer, 60.

### Language/Question Ambiguity Error Analysis

**Analysis:** There is an ambiguity in interpreting the phrase "2019 game that won the British Academy Games Awards." One interpretation refers to the game that won the "Best Game" award at the 2019 BAFTA Games Awards (British Academy Games Awards), regardless of when it was actually released. The other interpretation requires the game to have been both released in 2019 and won the BAFTA Game Award. This difference leads to inconsistent results, as the agent prefers the former (based solely on the award year), while the ground truth requires both conditions (release year and award) to be met.

#### Agent Content:

{ "subtask": "Search for the **2019 game that won the British Academy Games Awards (BAFTA)** and identify the specific game.",

"result": "2019 BAFTA Games Awards Winner God of War Based on the information from the 15th British Academy Games Awards held on April 4, 2019, **God of War** (2018) was the winner of the Best Game award. Game Details - **Developer**: Santa Monica Studio - **Publisher**: Sony Interactive Entertainment Europe - **Platform**: PlayStation 4 - **Release Date**: April 2018 Awards Won (5 total) 1. Best Game 2. Audio Achievement 3. Music 4. Narrative 5. Performer (Jeremy Davies for his role as "The Stranger") Additional Information - God of War received 10 nominations in total, making it the most nominated game at the ceremony - The 15th BAFTA Games Awards honored games released in 2018, though the ceremony was held in 2019" }

### Context Exceed: Error Case Level3 #5

**Question:** What animals that were mentioned in both Ilias Lagkouvardos's and Olga Tapia's papers on the alvei species of the genus named for Copenhagen outside the bibliographies were also present in the 2021 article cited on the alvei species' Wikipedia page about a multicenter, randomized, double-blind study?

**Answer:** mice

**Annotation steps:** 1. Searched "alvei copenhagen" on Google. 2. Opened [https://en.wikipedia.org/wiki/Hafnia\\_\(bacterium\)](https://en.wikipedia.org/wiki/Hafnia_(bacterium)). 3. Searched "Ilias Lagkouvardos hafnia alvei" on Google. 4. Opened [https://www.mdpi.com/2076-2607/11/1/123?type=check\\_update&version=2](https://www.mdpi.com/2076-2607/11/1/123?type=check_update&version=2). 5. Opened a new tab. 6. Searched "Olga Tapia hafnia alvei" on Google. 7. Opened <https://pubmed.ncbi.nlm.nih.gov/36080356/>. 8. Found all animals mentioned in the first paper. 9. Searched each animal from the first paper in the second paper. 10. Noted the animals mentioned in both outside the bibliographies. 11. Went back to the Wikipedia article. 12. Opened the link in the references to "The Probiotic Strain H. alvei HA4597® Improves Weight Loss in Overweight Subjects under Moderate Hypocaloric Diet: A Proof-of-Concept, Multicenter Randomized, Double-Blind Placebo-Controlled Study". 13. Opened the PDF. 14. Found the animals shared by all three papers.

### Context Exceed Error Analysis

**Analysis:** The content of Olga Tapia's research has been extracted correctly, but the mice mentioned in the text have not been extracted from the content successfully.

**Agent Content:**

```
{ "role": "assistant",  
  "tool_calls": [ {  
    "type": "function",  
    "function": {  
      "name": "extract_document_content",  
      "arguments": "{ \"document_path\": \"https://www.mdpi.com/1420-3049/27/17/5585\", \"query\": \"Olga Tapia research on Hafnia alvei with mentions of animals\" } } } ] }
```

### Limited Coding Capability: Error Case Level1 #36

**Question:** Bob was invited to participate in a game show, and he advanced to the final round. The final round offered Bob the chance to win a large sum by playing a game against the host. The host has 30 shiny prop coins, each of which is worth \$1,000 if Bob manages to win them by playing the game. The host hides the coins in three different prize boxes and then shuffles their order. The only rule restricting the host's coin placement is that one box must contain at least 2 coins, and one box must contain 6 more coins than another box. In order to play, Bob must submit three guesses, one guess for the number of coins in each box. The box is then opened and the number of coins is revealed. If Bob's guess is a number greater than the number of coins in the box, Bob earns no coins. If Bob guesses a number equal to or less than the number of coins in the box, Bob wins a number of coins equal to his guess. If Bob plays uses the optimal strategy, what's the minimum amount of money he can win from the game?

**Answer:** 16000

**Annotation steps:** Step 1: Evaluate the problem statement provided by my user, storing the relevant information: 30 coins with a value of \$1,000 distributed between 3 boxes. Each box must contain at least 2 coins. One box must contain 6 more coins than another. Step 2: Evaluate the base distribution: 2-8-20, noting that two boxes must contain at least 8 coins. Step 3: Evaluate the most even allowable distribution: 8,8,14, noting that two boxes must contain at least 8 coins. Step 4: Evaluate a case where Bob guesses 8 for each box in the outlier distributions. Step 5: For the worst case 2-8-20 distribution, Bob wins  $0+8+8 = 16$  coins. Step 6: For the 8-8-14 distribution, Bob wins  $8+8+8 = 24$  coins. Step 7: Convert the worst-case coin count to a prize value,  $16 * \$1,000 = \$16,000$ . Step 8: Report the correct answer to my user: "\$16,000"

### Limited Coding Capability Error Analysis

**Analysis:** The agent ignored the hard constraint that each box must contain at least 2 coins when writing the code, and mistakenly considered the distribution of 0 and 1 coins (such as (0,6,24), (1,7,22), etc.) as legal.

#### Agent Content:

{ "subtask": "Implement a solution using code to systematically explore all valid coin distributions, calculate Bob's optimal guesses for each distribution, and determine the minimum guaranteed winnings.",

"result": "Optimal Strategy for the Game Show Problem I've implemented a solution to systematically explore all valid coin distributions, calculate Bob's optimal guesses for each distribution, and determine the minimum guaranteed winnings. Let me explain my approach and findings. Solution Approach 1. **\*\*Generate Valid Distributions\*\***: I created a function to generate all valid distributions of 30 coins among 3 boxes that satisfy the constraints: - Total of 30 coins - **One box must contain at least 2 coins** - One box must contain 6 more coins than another box 2. **\*\*Calculate Winnings\*\***: For each distribution and guessing strategy, I calculated the maximum possible winnings by trying all permutations of the guesses (since Bob can choose which guess to apply to which box). 3. **\*\*Find Optimal Strategy\*\***: I implemented a systematic search to find the strategy that maximizes the minimum guaranteed winnings across all possible distributions. Key Findings 1. **\*\*Valid Distributions\*\***: The code identified 12 valid distributions of 30 coins that satisfy all constraints. 2. **\*\*Optimal Strategy\*\***: **The optimal strategy for Bob is to guess **\*\*[0, 6, 12]\*\*** for the three boxes.** 3. **\*\*Minimum Guaranteed Winnings\*\***: With this strategy, Bob is guaranteed to win at least **\*\*\$18,000\*\***, regardless of how the host distributes and shuffles the coins. 4. **\*\*Worst Case Scenario\*\***: The worst-case distribution for Bob is **\*\* (0, 6, 24) \*\***, where one box has 0 coins, one has 6 coins, and one has 24 coins. 5. **\*\*Verification\*\***: I verified that with the optimal strategy against the worst-case distribution, Bob indeed wins exactly \$18,000. Comparison with Other Strategies I also tested several other strategies mentioned in the dependency results: - Strategy (6, 6, 6): Minimum guaranteed \$12,000 - Strategy (2, 6, 10): Minimum guaranteed \$16,000 - Strategy (1, 5, 12): Minimum guaranteed \$17,000 - Strategy (7, 7, 7): Minimum guaranteed \$7,000 - Strategy (8, 8, 8): Minimum guaranteed \$8,000 - Strategy (6, 6, 7): Minimum guaranteed \$13,000 - Strategy (6, 7, 7): Minimum guaranteed \$13,000 This confirms that [0, 6, 12] is indeed the optimal strategy, guaranteeing \$18,000 in winnings. How the

Optimal Strategy Works With the  $[0, 6, 12]$  strategy: 1. Bob guesses 0 for the box with the fewest coins 2. Bob guesses 6 for the box with the middle number of coins 3. Bob guesses 12 for the box with the most coins This strategy works because: - It ensures Bob always wins from at least two boxes - Even in the worst-case distribution (0, 6, 24), Bob can match his guesses optimally to win \$18,000 - No other strategy can guarantee more than \$18,000 across all possible distributions Therefore, \$18,000 is the minimum amount Bob can win using the optimal strategy.” }