

Supplementary Materials

Cantor^{✍️}: Inspiring Multimodal Chain-of-Thought of MLLM

Anonymous Authors

1 PROMPTS USED IN CANTOR

Cantor consists of two stages: Decision-Generation and Execution. For different models, Cantor (Gemini) and Cantor (GPT-3.5) use the same prompt during the Decision-Generation stage and the Execute-Synthesis stage. During the Execute-Synthesis stage, we use the generated sub-tasks as the prompts. In terms of prompts for different datasets, only the in-context learning examples are different, while the other prompts are used the same.

1.1 Decision-Generation

In the Decision-Generation stage, the prompt template we used includes task instructions and in-context learning examples. The task instructions, as shown in Fig 1, actually include the guidance for the task, the functional definition of the expert module, and the format requirements for the answers. We design different in-context learning examples based on the characteristics of ScienceQA and MathVista. For ScienceQA, we use in-context learning examples as shown in Figs 2, 3, and 4. For MathVista, we use in-context learning examples as shown in Figs 5 and 6.

1.2 Execute-Modularization

In the Execute-Modularization stage, Cantor executes the sub-tasks assigned by the Decision-Generation by calling the expert module played by MLLM. Therefore, at this stage, we do not use manually designed prompts and directly input the assigned sub-tasks into MLLM as prompts, in the specific format of [Expert Module: a corresponding sub-task]. For example: [ChartSense Expert: Extract the values of all the bars from the chart.] and [VisionIQ Analyst: What is the total number of people in the image?].

1.3 Execute-Synthesis

We use the prompt template shown in Fig 7 to generate answers during the Execute-Synthesis stage. This includes prompts for generating answers and formatting requirements for answers. This includes three key points: 1. Play the role of an Answer Generator who is knowledgeable and adept at integrating information. 2. Think carefully before answering. 3. Maintain rationality and criticality when dealing with supplementary information.

2 ADDITIONAL ANALYSIS OF CANTOR

2.1 Case Presentation

In Figs 8, 9, 10, we show some specific cases of Cantor. It can be seen that Cantor has good decision-generating and practical problem-solving abilities.

2.2 Ablation Study with Modules

In this section, we further analyze the ablation study of the expert module. In the ablation experiment in the main text, we find that enabling a module has a greater impact on model performance than

disabling a module. We speculate that this is because when MLLM acts as various experts, it possesses a certain degree of universal higher-level information capture capability. As shown in Fig 11, after disabling ChartSense Expert, Cantor will adaptively adjust the decision and instead ask VisionIQ Analyst questions to obtain information about the chart. And VisionIQ Analyst also correctly answers this sub-task and facilitates the final inference to obtain the correct answer. This case illustrates that thanks to the versatility of MLLM, when playing various experts using MLLM, even if one expert module is disabled, the remaining expert modules can to some extent compensate for the lack of that module.

However, disabling a certain expert module still affects the integrity of Cantor. As shown in Fig 12, for chart information extraction, compared to ChartSense Expert, when using VisionIQ Analyst, it only extracts data for three bars and ignores the other bars with values of 0. This indicates that although different expert modules are to some extent universal, they are not omnipotent. Specific expert modules still focus on specific abilities and are lacking in other professional abilities. This also demonstrates the importance and irreplaceability of the four expert modules we propose. At the same time, we believe that thanks to the excellent scalability of Cantor, introducing more expert modules with different functions will further improve its performance.

2.3 Impact of Visual Information Levels

In this section, we demonstrate the impact of different levels of visual information on Gemini’s decision generation. As shown in Fig 13, when asking which country is highlighted, the model cannot answer the question both in the absence of visual information and with only a rough caption provided. This is because the model cannot acquire effective visual information solely from the question or a rough caption. When entering a detailed caption, even if it contains a lot of content, it is irrelevant information and lacks key information about what the highlighted country is. The model still cannot answer the question. Only by inputting images can the model obtain sufficient visual information for problem-solving.

As shown in Fig 14, another case is shown. When the detailed caption contains key information to answer the question, the model can also provide the correct answer. However, it should be noted that in practical applications, we cannot control whether the captions include key information for solving problems. On the contrary, the image must contain clues to the problem-solving. Therefore, inputting images is the best way to obtain visual information during decision generation.

Decision Generation Prompt

You are an advanced question-answering agent equipped with four specialized modules to aid in analyzing and responding to queries about images:

1. TextIntel Extractor: This module extracts and converts text within images into editable text format. It's particularly useful for images containing a mix of text and graphical elements. When this module is required, specify your request as: "TextIntel Extractor: <specific task or information to extract>."

2. ObjectQuant Locator: This module identifies and locates objects within an image. It's adept at counting objects and determining their spatial arrangement. When you need this module, frame your request as: "ObjectQuant Locator: <object1, object2, ..., objectN>," listing the objects you believe need detection for further analysis.

3. VisionIQ Analyst: This module processes and interprets visual data, enabling you to ask any queries related to the image's content. When information from this module is needed, phrase your request as: "VisionIQ Analyst: <your question about the image>."

4. ChartSense Expert: This module specializes in analyzing and interpreting information from charts and graphs. It can extract data points, understand trends, and identify key components such as titles, axes, labels, and legends within a chart. When you require insights from a chart or graph, specify your request as: "ChartSense Expert: <specific aspect of the chart you're interested in or question you have about the chart>."

When faced with a question about an image, which will be accompanied by a hint that might not cover all its details, your task is to:

If the question can be answered directly based on the information provided without the need for detailed input from the modules, specify this explicitly. Do not disclose the answer itself.

Otherwise:

- Provide a rationale for your approach to answering the question, explaining how you will use the information from the image and the modules to form a comprehensive answer.
- Assign specific tasks to each module as needed, based on their capabilities, to gather additional information essential for answering the question accurately.

Your response should be structured as follows:

Answer:

["This question does not require any modules and can be answered directly based on the information provided."] or [Rationale: Your explanation of how you plan to approach the question, including any initial insights based on the question and image information provided. Explain how the modules' input will complement this information.]

Modules' tasks (if applicable):

1. TextIntel Extractor: [Specify the text or information to be extracted from the image, if necessary.]
2. ObjectQuant Locator: [List the objects to be identified or counted in the image, if required.]
3. VisionIQ Analyst: [Pose any specific questions you have about the image that require deeper visual analysis, if applicable.]
4. ChartSense Expert: [Extract chart data or specify any questions about the chart, if required.]

Ensure your response adheres to this format to systematically address the question using the available modules or direct analysis as appropriate.

Here are some examples:

...

Please refer to the prompts and examples above to help me solve the following problem: <Question>.

Figure 1: The prompt of the Decision-Generation stage.



In-context Learning Examples (ScienceQA)

Here are some examples:

"Question1": "Which solution has a higher concentration of blue particles?",
"Choices": [
"Solution B",
"neither; their concentrations are the same",
"Solution A"
],

Answer:
1. Concentration in a solution refers to the amount of a substance (solute) present in a specified amount of another substance (solvent).
2. To know the solvent volume, we need TextIntel Extractor to extract information about the volume from the image.
3. To understand the number of blue particles in solution A and solution B, we need a ObjectQuant Locator to detect them.

Modules' tasks:
1. TextIntel Extractor: Extract keywords related to solution volume in Solution A and Solution B.
2. ObjectQuant Locator: the number of blue particles in Solution A and Solution B.

"Question2": "Compare the average kinetic energies of the particles in each sample. Which sample has the higher temperature?",
"Choices": [
"neither; the samples have the same temperature",
"sample B",
"sample A"
],

Answer:
1. The temperature of particles in a substance is directly proportional to its average kinetic energy. The formula for kinetic energy of an particle is $\frac{1}{2}mv^2$, where m represents the mass of the object and v represents its velocity.
2. So TextIntel Extractor is needed to retrieve key information about mass and speed in the image.

Modules' tasks:
1. TextIntel Extractor: Extract keywords related to mass and speed in sample A and sample B.

Figure 2: In-context Learning Examples on ScienceQA.

In-context Learning Examples (ScienceQA)

"Question3": "Think about the magnetic force between the magnets in each pair. Which of the following statements is true?",

"Choices": [

"The magnetic force is stronger in Pair 2.",

"The magnetic force is stronger in Pair 1.",

"The strength of the magnetic force is the same in both pairs."

],

Answer:

1. The magnitude of magnetic force is independent of the direction of the magnetic pole, and is not related to whether the magnetic pole is attracted or repelled.
2. Different pairs of magnets will not affect each other.
3. The magnitude of magnetic force is inversely proportional to the distance between two magnets.
4. To know the distance between magnets, TextIntel Extractor is needed to extract distance information.

Modules' tasks:

1. TextIntel Extractor: Extract the distance values between two pairs of magnets in pair1 and pair2 separately.

"Question4": "What is the expected ratio of offspring with mutated antennae to offspring with normal antennae? Choose the most likely ratio.",

"Choices": [

"1:3",

"0:4",

"3:1",

"2:2",

"4:0"

]

Answer:

1. In a group of fruit flies, some individuals have mutated antennae and others have normal antennae. In this group, the gene for the antenna type trait has two alleles.
2. The allele 'A' is for mutated antennae, the allele 'a' is for normal antennae.
3. To know the specific gene composition, we need TextIntel Extractor to extract information from Punnett square.

Modules' tasks:

1. TextIntel Extractor: Extract the genotypes from the Punnett square.

Figure 3: In-context Learning Examples on ScienceQA.

In-context Learning Examples (ScienceQA)

"Question5": "Which month is the wettest on average in Christchurch?",

"Choices": [

"August",

"April",

"May"

]

Answer:

1. This question provides a chart of Christchurch precipitation. We need to compare the monthly precipitation in Christchurch to determine which month is the wettest.

2. In order to determine which month has the highest precipitation, we need ChartSense Expert to extract the precipitation for each month

Modules' tasks:

1. ChartSense Expert: Extract the precipitation for each month from the chart.

"Question6": "What is the capital of New Jersey?",

"Choices": [

"Augusta",

"Montpelier",

"Newark",

"Trenton"

]

Answer:

1. This question does not require any modules and can be answered directly based on the information provided.

Figure 4: In-context Learning Examples on ScienceQA.

In-context Learning Examples (MathVista)

Here are some examples:

"Question1": Subtract all large rubber spheres. Subtract all big shiny cylinders. How many objects are left?

Answer:

1. First, we need to know the quantity of all objects. Secondly, it is important to know the number of large rubber spheres and the number of big shiny cylinders. Then perform the subtraction operation again.
2. To know the number of all objects, we need a ObjectQuant Locator.
3. To know the number of large rubber spheres, we need a ObjectQuant Locator.
4. To know the number of big shiny cylinders, we need a ObjectQuant Locator.

Modules' tasks:

1. ObjectQuant Locator: the number of all objects.
2. ObjectQuant Locator: the number of large rubber spheres.
3. ObjectQuant Locator: the number of big shiny cylinders.

"Question2": "What is the age gap between these two people in image?"

Answer:

1. To determine the age gap between two individuals, it is first necessary to know who they are and their date of birth.
2. Calculate the difference in their birth dates to obtain the age difference.

Modules' tasks:

1. VisionIQ Analyst: Who are the two people in the picture and what is their date of birth?

"Question3": "What time is shown? Answer by typing a time word, not a number. It is () after eight.",

"choices": [

"half",

"quarter",

"o'clock",

"quarter to",

"quarter past"

]

Answer:

1. To know the specific time, it is necessary to determine which number the hour and minute of the clock point to respectively.
2. Combine the numbers pointed by the hour hand and minute hand to obtain the final time and fill in the blanks.

Modules' tasks:

1. VisionIQ Analyst: the number the hour hand is pointing to.
2. VisionIQ Analyst: the number the minute hand is pointing to.

Figure 5: In-context Learning Examples on MathVista.

In-context Learning Examples (MathVista)

"Question4": "Is Light Seafoam less than Dark Salmon?",

"choices": [

"yes",

"no"

]

Answer:

1. We need to know the numerical values corresponding to Light Seafoam and Dark Salmon to determine their quantity relationship.
2. To obtain the numerical values corresponding to Light Seafoam and Dark Salmon, we need ChartSense Expert.

Modules' tasks:

1. ChartSense Expert: the value of Light Seafoam.
2. ChartSense Expert: the value of Dark Salmon.

"Question5": "How many algorithms have accuracies higher than 9?",

Answer:

1. First, we need to know the accuracy of each algorithm.
2. Then we can compare the accuracy of each algorithm with 9.
3. Finally, we can count the number of algorithms that have accuracy higher than 9.
4. To know the accuracy of each algorithm, we need ChartSense Expert to extract the information from the chart.

Modules' tasks:

1. ChartSense Expert: Extract the accuracy of each algorithm from the chart.

Figure 6: In-context learning examples on MathVista.

Execute-Synthesis Prompt

You are a knowledgeable and skilled information integration science expert. Please gradually think and answer the questions based on the given questions, options, and supplementary information. Please note that we not only need answers, but more importantly, we need rationales for obtaining answers. Please combine your knowledge and supplementary information to obtain reasoning and answers. Please prioritize using your knowledge to answer questions. If unable to answer, maintain critical thinking and select effective information to assist you in selecting the most correct option as the answer. Furthermore, please do not rely solely on supplementary information, as the provided supplementary information may not always be effective. Please do not answer with uncertainty, try your best to give an answer.

The expected response format is as follows: Rationale:<rationale> Answer:<answer>.

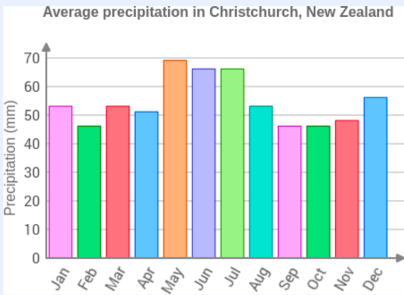
Please answer the following case: < Question> , <Supplementary information>.

Figure 7: The prompt of the Execute-Synthesis stage.

The Specific Case of Cantor

Which month is the wettest on average in Christchurch?

- Choices:
- (A) August
 - (B) April
 - (C) May



Supplementary Information:

1. This question provides a chart of Christchurch precipitation. We need to compare the monthly precipitation in Christchurch to determine which month is the wettest.

2. In order to determine which month has the highest precipitation, we need ChartSense Expert to extract the precipitation for each month.

Modules' tasks:

ChartSense Expert: Extract the precipitation for each month from the chart.

Answer: The precipitation for each month is:\n- January: 52 mm\n- February: 46 mm\n- March: 51 mm\n- April: 50 mm\n- May: 68 mm\n- June: 67 mm\n- July: 67 mm\n- August: 52 mm\n- September: 47 mm\n- October: 48 mm\n- November: 49 mm\n- December: 54 mm.

Answer Generator:

Rationale: The month with the highest precipitation is May, with 68 mm of rain.

Answer: May.

Will these magnets attract or repel each other?

- Choices:
- (A) repel
 - (B) attract



Supplementary Information:

1. Like poles of a magnet repel each other. Unlike poles of a magnet attract each other.

2. To know the poles of the magnets, TextIntel Extractor is needed to extract information about the poles.

Modules' tasks:

TextIntel Extractor: Extract the poles of the two magnets.

Answer: The poles of the magnets are: - Magnet 1: North (N) and South (S) - Magnet 2: South (S) and North (N).

Answer Generator:

Rationale: The south pole of magnet 1 is facing the south pole of magnet 2. Like poles repel each other.

Answer: repel.

Figure 8: The Specific Case of Cantor.



The Specific Case of Cantor

What is the probability that a cow produced by this cross will have solid coloring?

- Choices:
- (A) 4/4
 - (B) 0/4
 - (C) 1/4
 - (D) 2/4
 - (E) 3/4

	a	a
a	aa	aa
a	aa	aa



Cantor

Supplementary Information:

- 1. The probability of getting a cow with solid coloring is determined by the number of offspring with the genotype AA or Aa.
- 2. To know the specific gene composition, we need TextIntel Extractor to extract information from Punnett square.

Modules' tasks:

TextIntel Extractor: Extract the genotypes from the Punnett square.

Answer: aa, aa, aa, aa

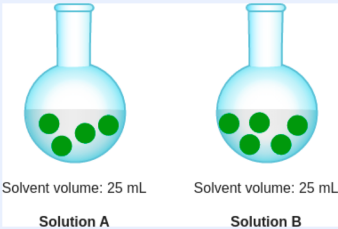
Answer Generator:

Rationale: Since all of the offspring have the aa genotype, this means that all of them will have white spots.

Answer: 0/4.

Which solution has a higher concentration of green particles?

- Choices:
- (A) Solution B
 - (B) neither; their concentrations are the same
 - (C) Solution A



Cantor

Supplementary Information:

- 1. Concentration in a solution refers to the amount of a substance (solute) present in a specified amount of another substance (solvent).
- 2. To know the solvent volume, we need TextIntel Extractor to extract information about the volume from the image.
- 3. To understand the number of green particles in solution A and solution B, we need ObjectQuant Locator to detect them.

Modules' tasks:

TextIntel Extractor: Extract keywords related to solution volume in Solution A and Solution B.

Answer: Solution A: 25 mL Solution B: 25 mL

ObjectQuant Locator: the number of green particles in Solution A and Solution B.

Answer: Solution A has fewer green particles than Solution B.

Answer Generator:

Rationale: Solution B has more green particles than Solution A.

Answer: Solution B.

Figure 9: The Specific Case of Cantor.

The Specific Case of Cantor

What is the perimeter of the rectangle?



Supplementary Information:

To determine the perimeter of the rectangle, we need to know the length of all four sides and then add them together. Since the opposite sides of a rectangle are equal, we can calculate the perimeter using the formula: $\text{Perimeter} = 2(\text{length} + \text{width})$.

Modules' tasks:

VisionIQ Analyst: Determine the length and width of the rectangle.

Answer: In this case, VisionIQ Analyst would determine that the length of the rectangle is 2 yards and the width is 1 yard.

Answer Generator:

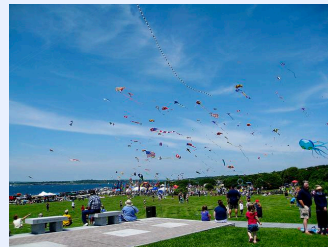
Rationale: To determine the perimeter of the rectangle, we need to know the length of all four sides and then add them together. Since the opposite sides of a rectangle are equal, we can calculate the perimeter using the formula: $\text{Perimeter} = 2(\text{length} + \text{width}) = 2(2+1) = 6$.

Answer: 6 yd.

What is the blue kite in the lower right corner shaped like?

Choices:

- (A) ferre
- (B) cat
- (C) cloud
- (D) octopus



Supplementary Information:

- To determine the shape of the blue kite in the lower right corner, we need to identify the blue kite in the lower right corner and then determine its shape.
- To identify the blue kite in the lower right corner, we need to know the location of the blue kite in the lower right corner.
- To determine the shape of the blue kite in the lower right corner, we need to know the shape of the blue kite in the lower right corner.

Modules' tasks:

ObjectQuant Locator: Identify the location of the blue kite in the lower right corner.

Answer: The blue kite in the lower right corner is located in the sky above the people.

VisionIQ Analyst: Determine the shape of the blue kite in the lower right corner.

Answer: The blue kite in the lower right corner is an octopus.

Answer Generator:

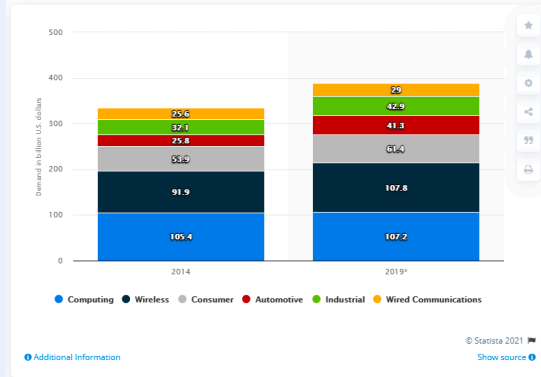
Rationale: The blue kite in the lower right corner is shaped like an octopus because it has eight tentacles.

Answer: (D)

Figure 10: The Specific Case of Cantor.

Ablation Study with Expert Module

What's the computing and wireless total for semiconductor demand in 2014?



Cantor

Supplementary Information:

1. First, we need to know the computing demand and wireless demand in 2014. Then we can add them together to get the total.
2. To know the computing demand and wireless demand in 2014, we need to extract the information from the chart.

Modules' tasks:

ChartSense Expert: Extract the computing demand in 2014 from the chart.

Answer: The computing demand in 2014 was 105.4 billion dollars.

ChartSense Expert: Extract the wireless demand in 2014 from the chart.

Answer: 91.9 billion dollars.

Answer Generator:

Rationale: First, we need to know the computing demand and wireless demand in 2014. Then we can add them together to get the total. To know the computing demand and wireless demand in 2014, we need to extract the information from the chart. The computing demand in 2014 was 105.4 billion dollars. The wireless demand in 2014 was 91.9 billion dollars. So, the computing and wireless total for semiconductor demand in 2014 is $105.4 + 91.9 = 197.3$ billion dollars.

Answer: 197.3.



Cantor (without ChartSense Expert)

Supplementary Information:

1. First, we need to know the computing and wireless demand in 2014. Then we can add these two numbers together to get the total.
2. To know the computing demand in 2014, we need the VisionIQ Analyst to extract the information from the chart.
3. To know the wireless demand in 2014, we need the VisionIQ Analyst to extract the information from the chart.

Modules' tasks:

VisionIQ Analyst: What's the computing demand in 2014?

Answer: The computing demand for VisionIQ Analyst in 2014 was 105.4 billion US dollars.

VisionIQ Analyst: What's the wireless demand in 2014?

Answer: The wireless demand in 2014 was \$91.9 billion.

Answer Generator:

Rationale: First, we need to know the computing and wireless demand in 2014. Then we can add these two numbers together to get the total. The computing demand in 2014 was 105.4 billion US dollars. The wireless demand in 2014 was \$91.9 billion. So, the computing and wireless total for semiconductor demand in 2014 is $105.4 + 91.9 = 197.3$ billion US dollars.

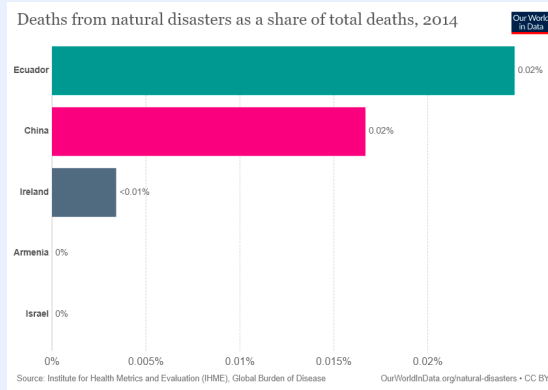
Answer: 197.3.

Figure 11: Case study on enabling and disabling Expert Module.

Ablation Study with Expert Module

Is the sum of two lowest bar is greater then the largest bar?

- (A) Yes
(B) No



Cantor

Supplementary Information:

1. First, we need to know the values of all the bars.
2. Then we can find the two lowest bars and add their values.
3. Finally, we can compare the sum of the two lowest bars with the value of the largest bar.
4. To know the values of all the bars, we need ChartSense Expert to extract the information from the chart.

Modules' tasks:

ChartSense Expert: Extract the values of all the bars from the chart.

Answer: Ecuador: 0.02% China: 0.02% Ireland: <0.01% Armenia: 0% Israel: 0%.

Answer Generator:

Rationale: First, we need to know the values of all the bars. We can use the ChartSense Expert to extract the information from the chart. The values of the bars are: Ecuador: 0.02%, China: 0.02%, Ireland: <0.01%, Armenia: 0%, Israel: 0%. Then we can find the two lowest bars and add their values. The two lowest bars are Armenia and Israel, and their values are 0% and 0%, respectively. So, the sum of the two lowest bars is $0\% + 0\% = 0\%$. Finally, we can compare the sum of the two lowest bars with the value of the largest bar. The largest bar is Ecuador, and its value is 0.02%. So, the sum of the two lowest bars is 0%, which is less than the value of the largest bar, 0.02%. Therefore, the answer is (B) No.

Answer: (B).

Cantor (without ChartSense Expert)

Supplementary Information:

1. To solve this problem, we need to know the values of all the bars.
2. Then we can calculate the sum of the two lowest bars and compare it with the largest bar.
3. Finally, we can determine whether the sum of the two lowest bars is greater than the largest bar.
4. To know the values of all the bars, we need VisionIQ Analyst to extract the information from the chart.

Modules' tasks:

VisionIQ Analyst: Extract the values of all the bars from the chart.

Answer: Ecuador: 0.02%\nChina: 0.02%\nIreland: <0.01%

Answer Generator:

Rationale: The sum of two lowest bar is 0.02% (China) + 0.01% (Ireland) = 0.03% . The largest bar is 0.02% (Ecuador). $0.03\% > 0.02\%$. So the answer is yes. Answer: (A).

Figure 12: Case study on enabling and disabling Expert Module.

