

## Supplementary Material for SCIBENCH

<b>A SciBench: Textbook Sources</b>	<b>13</b>
A.1 Textbook . . . . .	13
A.2 Examination . . . . .	14
A.3 Textbook Examples . . . . .	14
<b>B SciBench: More Statistics</b>	<b>14</b>
B.1 UI Design . . . . .	14
<b>C Experimental Details</b>	<b>14</b>
C.1 Prompting . . . . .	14
C.2 Experiment Process . . . . .	17
<b>D Problem Solving Abilities of Current LLMs</b>	<b>18</b>
D.1 Example . . . . .	18
D.2 Assessment of evaluation protocol . . . . .	19

### A SCIBENCH: TEXTBOOK SOURCES

#### A.1 TEXTBOOK

- PHYSICAL CHEMISTRY, ATKINS ET AL. [40] (atkins) provides an exploration of equilibrium, structure, and reactions, integrating contemporary techniques like nanoscience, spectroscopy, and computational chemistry.
- QUANTUM CHEMISTRY, MCQUARRIE [39] (chemmc) meticulously covers Quantum Mechanics, from foundational principles like blackbody radiation and Heisenberg’s Uncertainty Principle to complex topics such as Schrödinger’s equation, quantum mechanical operators, and the application of quantum mechanics in chemical bonding.
- QUANTUM CHEMISTRY, LEVINE ET AL. [38] (quan) explores quantum chemistry, providing a detailed understanding of the Schrödinger equation, particle behavior in various scenarios, quantum mechanics operators, and other foundational quantum principles. It delves into specific applications like the electronic structure of diatomic and polyatomic molecules, variation methods, perturbation theory, electron spin and its implications in quantum mechanics, as well as various computational methods for molecular quantum mechanics.
- PHYSICAL CHEMISTRY, QUANTA, MATTER, AND CHANGE, ATKINS ET AL. [25] (matter) combines physics and mathematics, beginning with basics like differentiation and integration, advancing through quantum mechanics and atomic structure, then exploring thermodynamics, molecular motion, and chemical kinetics. Each section is supplemented with mathematical concepts such as differential equations, vectors, and probability theory.
- CLASSICAL DYNAMICS OF PARTIAL AND SYSTEMS, THORNTON AND MARION [37] (class) initiates with an exploration of fundamental mathematical concepts, discussing scalars, vectors, matrix operations, coordinate transformations, differentiation, and integration of vectors, using these constructs to illustrate concepts like velocity, acceleration, and angular velocity. It then transitions into the realm of Newtonian mechanics, detailing Newton’s laws, frames of reference, and the equation of motion for a single particle.
- THERMODYNAMICS, STATISTICAL THERMODYNAMICS, AND KINETICS, [36] (thermo) navigates through thermodynamics’ principles, from fundamental concepts to complex laws, further discussing real and ideal gases, solutions, electrochemical cells, and statistical thermodynamics. It concludes with an examination of the kinetic theory of gases, transport phenomena, and chemical kinetics.

- FUNDAMENTALS OF PHYSICS, HALLIDAY ET AL. [35] (fund) covers undergraduate physics topics, ranging from fundamental concepts like motion and energy to more advanced areas such as quantum physics and nuclear physics.
- ELEMENTARY DIFFERENTIAL EQUATIONS AND BOUNDARY VALUE PROBLEMS, [43] (diff) provides a detailed exploration of differential equations, progressing from basic mathematical models to advanced topics like the Laplace Transform, linear systems, numerical methods, and Fourier series. It culminates with a deep dive into nonlinear equations, partial differential equations, and boundary value problems.
- PROBABILITY AND STATISTICAL INFERENCE, [42] (stat) covers probability and statistics, including fundamental concepts, discrete and continuous distributions, bivariate distributions, functions of random variables, and estimation techniques.
- CALCULUS: EARLY TRANSCENDENTALS, [41] (calculus) begins with diagnostic tests in foundational topics, and explores functions from multiple perspectives. It comprehensively covers calculus concepts from limits to three-dimensional analytic geometry, incorporating applications in various fields.

## A.2 EXAMINATION

- INTRODUCTION TO DATA MINING provides an introductory survey of data mining, which involves the automatic discovery of patterns, associations, changes, and anomalies in large databases. It explores various application areas of data mining, including bioinformatics, e-commerce, environmental studies, financial markets, multimedia data processing, network monitoring, and social service analysis.
- FUNDAMENTALS ARTIFICIAL INTELLIGENCE provides an introduction to the core problem-solving and knowledge representation paradigms in artificial intelligence. It covers Lisp programming with regular assignments, as well as topics such as search methods, planning techniques, knowledge structures, natural language processing, expert systems, vision, and parallel architectures.
- DIFFERENTIAL EQUATIONS covers various topics in differential equations, including first-order and second-order linear equations with constant coefficients, power series solutions, and linear systems. Students will explore the principles and applications of these mathematical concepts.

## A.3 TEXTBOOK EXAMPLES

# B SCIBENCH: MORE STATISTICS

## B.1 UI DESIGN

We employed a team of seven individuals to gather data from textbooks using an annotation tool. Each individual was responsible for 1-2 books, encompassing approximately 100 examples. The user interface of the annotation tool is depicted in Figure S2. For subsequent verification, we preserved images of problems and their corresponding answers. To ensure clarity in future references, we have maintained the original sequence of problems as they appear in the textbooks.

# C EXPERIMENTAL DETAILS

## C.1 PROMPTING

ChatGPT and GPT-4’s API have three message parameters: SYSTEM, USER, and ASSISTANT. The SYSTEM parameter represents the system prompt, which provides context and instructions to the model. The USER parameter is the training prompt or input provided by the user, and the ASSISTANT parameter contains the model’s output or response. We provide all system prompts and training prompts used in our experiments as below.

### **System Prompt for Zero-Shot, Few-Shot, and Chain-of-Thought setting:**

Please provide a clear and step-by-step solution for a scientific problem in the categories of

<p><b>Problem (fund)</b> Two charged particles are fixed to an <math>x</math> axis: Particle 1 of charge <math>q_1 = 2.1 \times 10^{-8} \text{ C}</math> is at position <math>x = 20 \text{ cm}</math> and particle 2 of charge <math>q_2 = -4.00q_1</math> is at position <math>x = 70 \text{ cm}</math>. At what coordinate on the axis (other than at infinity) is the net electric field produced by the two particles equal to zero? <b>Answer:</b> <math>-30 \text{ cm}</math></p>
<p><b>Problem (thermo)</b> <math>\text{N}_2\text{O}_3</math> dissociates according to the equilibrium <math>\text{N}_2\text{O}_3(\text{g}) \rightleftharpoons \text{NO}_2(\text{g}) + \text{NO}(\text{g})</math>. At 298 K and one bar pressure, the degree of dissociation defined as the ratio of moles of <math>\text{NO}_2(\text{g})</math> or <math>\text{NO}(\text{g})</math> to the moles of the reactant assuming no dissociation occurs is <math>3.5 \times 10^{-3}</math>. Calculate <math>\Delta G_R^\circ</math> for this reaction. <b>Answer:</b> <math>28 \text{ kJ mol}^{-1}</math></p>
<p><b>Problem (class)</b> Halley's comet, which passed around the sun early in 1986, moves in a highly elliptical orbit with an eccentricity of 0.967 and a period of 76 years. Calculate its minimum distances from the Sun. <b>Answer:</b> <math>8.8 \times 10^{10} \text{ m}</math></p>
<p><b>Problem (quan)</b> A one-particle, one-dimensional system has <math>\Psi = a^{-1/2} e^{- x /a}</math> at <math>t = 0</math>, where <math>a = 1.0000 \text{ nm}</math>. At <math>t = 0</math>, the particle's position is measured. Find the probability that the measured value is between <math>x = 0</math> and <math>x = 2 \text{ nm}</math>. <b>Answer:</b> 0.4908</p>
<p><b>Problem (chemmc)</b> One of the most powerful modern techniques for studying structure is neutron diffraction. This technique involves generating a collimated beam of neutrons at a particular temperature from a high-energy neutron source and is accomplished at several accelerator facilities around the world. If the speed of a neutron is given by <math>v_n = (3k_B T/m)^{1/2}</math>, where <math>m</math> is the mass of a neutron, then what temperature is needed so that the neutrons have a de Broglie wavelength of <math>50 \text{ pm}</math>? <b>Answer:</b> 2500 K</p>
<p><b>Problem (atkins)</b> The change in molar internal energy when <math>\text{CaCO}_3(\text{s})</math> as calcite converts to another form, aragonite, is <math>+0.21 \text{ kJ mol}^{-1}</math>. Calculate the difference between the molar enthalpy and internal energy changes when the pressure is 1.0 bar given that the densities of the polymorphs are <math>2.71 \text{ g cm}^{-3}</math> and <math>2.93 \text{ g cm}^{-3}</math>, respectively. <b>Answer:</b> <math>-0.28 \text{ Pa m}^3 \text{ mol}^{-1}</math></p>
<p><b>Problem (matter)</b> In an industrial process, nitrogen is heated to 500 K at a constant volume of <math>1.000 \text{ m}^3</math>. The gas enters the container at 300 K and 100 atm. The mass of the gas is 92.4 kg. Use the van der Waals equation to determine the approximate pressure of the gas at its working temperature of 500 K. For nitrogen, <math>a = 1.39 \text{ dm}^6 \text{ atm mol}^{-2}</math>, <math>b = 0.0391 \text{ dm}^3 \text{ mol}^{-1}</math>. <b>Answer:</b> 140 atm</p>
<p><b>Problem (calc)</b> A planning engineer for a new alum plant must present some estimates to his company regarding the capacity of a silo designed to contain bauxite ore until it is processed into alum. The ore resembles pink talcum powder and is poured from a conveyor at the top of the silo. The silo is a cylinder 100ft high with a radius of 200ft. The conveyor carries ore at a rate of <math>60,000\pi \text{ ft}^3/\text{h}</math> and the ore maintains a conical shape whose radius is 1.5 times its height. If, at a certain time <math>t</math>, the pile is 60ft high, how long will it take for the pile to reach the top of the silo? <b>Answer:</b> 9.8 h</p>
<p><b>Problem (stat)</b> In a study concerning a new treatment of a certain disease, two groups of 25 participants in each were followed for five years. Those in one group took the old treatment and those in the other took the new treatment. The theoretical dropout rate for an individual was 50% in both groups over that 5-year period. Let <math>X</math> be the number that dropped out in the first group and <math>Y</math> the number in the second group. Assuming independence where needed, give the sum that equals the probability that <math>Y \geq X + 2</math>. HINT: What is the distribution of <math>Y - X + 25</math>? <b>Answer:</b> 0.3359</p>
<p><b>Problem (diff)</b> Newton's law of cooling states that the temperature of an object changes at a rate proportional to the difference between its temperature and that of its surroundings. Suppose that the temperature of a cup of coffee obeys Newton's law of cooling. If the coffee has a temperature of <math>200^\circ \text{ F}</math> when freshly poured, and 1 min later has cooled to <math>190^\circ \text{ F}</math> in a room at <math>70^\circ \text{ F}</math>, determine when the coffee reaches a temperature of <math>150^\circ \text{ F}</math>. <b>Answer:</b> 6.07 min</p>

Figure S1: Textbook examples with acronym highlighted in brown.

Chemistry, Physics, or Mathematics. The problem will specify the unit of measurement, which should not be included in the answer. Express the final answer as a decimal number with three digits after the decimal point. Conclude the answer by stating "The answer is therefore  $\boxed{\text{ANSWER}}$ ."

#### System Prompt for Python setting:

Please provide a clear and step-by-step solution for a scientific problem in the categories of Chemistry, Physics, or Mathematics. The problem will specify the unit of measurement. Please translate the solution steps into Python code and encase the Python code within triple backticks for clarity.

#### System Prompt for Wolfram setting:

Please provide a clear and step-by-step solution for a scientific problem in the categories of Chemistry,

Welcome, SciBench! You are annotating #1 data.

**Whole problem in image format (Required)**

The logistic model has been applied to the natural growth of the halibut population in certain areas of the Pacific Ocean.<sup>12</sup> Let  $y$ , measured in kilograms, be the total mass, or biomass, of the halibut population at time  $t$ . The parameters in the logistic equation are estimated to have the values  $r = 0.71/\text{year}$  and  $K = 80.5 \times 10^6 \text{ kg}$ . If the initial biomass is  $y_0 = 0.25K$ , find the time  $\tau$  for which  $y(\tau) = 0.75K$ .

**Question part in image format (Optional)**

Click here or drag here your images for preview

**Answer part in image format (Required)**

**Problem Text**

The logistic model has been applied to the natural growth of the halibut population in certain areas of the Pacific Ocean.<sup>12</sup> Let  $y$ , measured in kilograms, be the total mass, or biomass, of the halibut population at time  $t$ . The parameters in the logistic equation are estimated to have the values  $r = 0.71/\text{year}$  and  $K = 80.5 \times 10^6 \text{ kg}$ . If the initial biomass is  $y_0 = 0.25K$ , find the time  $\tau$  for which  $y(\tau) = 0.75K$ .

**Solution**

To find  $\tau$ , we can first solve Eq. (12) for  $t$ . We obtain

$$e^{rt} = \frac{(y_0/K) \left[ 1 - (y/K) \right]}{(y/K) \left[ 1 - (y_0/K) \right]}$$

hence

$$t = -\frac{1}{r} \ln \frac{(y_0/K) \left[ 1 - (y/K) \right]}{(y/K) \left[ 1 - (y_0/K) \right]}$$

Using the given values of  $r$  and  $y_0/K$  and setting  $y/K = 0.75$ , we find that

$$\tau = -\frac{1}{0.71} \ln \frac{(0.25)(0.25)}{(0.75)(0.75)} = \frac{1}{0.71} \ln 9 \approx 3.095 \text{ years}$$

**Answer (LaTeX)**

3.095

**Answer (Number)**

3.095

**Unit**

years

**Source (book title, url or file name)**

dtf

**Problem ID (e.g. Example ID)**

2.5.1

**Comment**

Submit

Figure S2: The UI design of data annotation.

Physics, or Mathematics. The problem will specify the unit of measurement. Please translate the solution steps into Wolfram code and encase the Wolfram Language code within triple backticks for clarity.

### System Prompt for Evaluation Protocol:

Examine the given problem, the correct solution, and the model's solution. Identify the reason for the error in the model's solution based on the following 10 categories:

1. Logical Decomposition and Analysis Skills: This ability involves decomposing the problem into smaller, manageable parts, and understanding the relationships between these parts.
2. Identification of Assumptions: This skill involves the AI's ability to recognize relevant and necessary assumptions in the problem.
3. Spatial Perception: This is important for understanding problems in areas such as physics and chemistry, where you need to visualize molecules, forces, fields, etc.
4. Causal Reasoning: This is the ability to understand cause and effect relationships.
5. Problem Deduction Skills: This pertains to the ability to infer and deduce potential solutions or underlying principles from the given information in a problem.
6. Abstract Reasoning: This skill involves the ability to understand complex concepts that can't be perceived physically, and to recognize patterns or relationships beyond concrete examples.
7. Scientific Literacy: This skill involves a comprehensive understanding of key scientific principles, terminology, and methodologies across a range of disciplines.
8. Code Conversion Skills: This denotes the ability to accurately translate solution steps into different programming languages, like Python or Wolfram, without syntax errors.
9. Logical Reasoning: This is the ability to make a reasoned argument and to identify fallacies or inconsistencies in an argument or set of data.
10. Calculation Skills: This involves the ability to accurately carry out mathematical operations and computations.

Conclude your final error reason category number within `\boxed{}`.

### Training Prompt for Zero-Shot Chain-of-Thought:

Stage 1:

**Input:** [input-question] Let's think step by step.

**Output:** <explanation>

Stage 2:

**Input:** [input-question] Let's think step by step. [explanation] + Therefore, the answer is:

**Output:** <answer>

### Training Prompt for Few-Shot:

**Input:**

Problem 1: [Question 1] The answer is `\boxed{[Answer 1]}`.

Problem 2: [Question 2] The answer is `\boxed{[Answer 2]}`.

...

Problem n: [Question n] The answer is `\boxed{[Answer n]}`.

Problem n+1: [Question n+1]

**Output:** The answer is `\boxed{<answer>}`.

### Training Prompt for Few-Shot Chain-of-Thought:

**Input:**

Problem 1: [Question 1] Explanation for Problem 1: [Explanation 1]. The answer is `\boxed{[Answer 1]}`.

Problem 2: [Question 2] Explanation for Problem 2: [Explanation 2]. The answer is `\boxed{[Answer 2]}`.

...

Problem n: [Question n] Explanation for Problem n: [Explanation n]. The answer is `\boxed{[Answer n]}`.

Problem n+1: [Question n+1]

**Output:** Explanaiton for Problem n+1: <explanation>. The answer is `\boxed{<answer>}`.

### Training Prompt for Few-Shot Python/Wolfram:

**Input:**

Problem 1: [Question 1] Explanation for Problem 1: [Explanation 1]. Python/Wolfram language for Problem 1: ````[Python/Wolfram code 1]````.

Problem 2: [Question 2] Explanation for Problem 2: [Explanation 2]. Python/Wolfram language for Problem 2: ````[Python/Wolfram code 2]````.

...

Problem n: [Question n] Explanation for Problem n: [Explanation n]. Python/Wolfram language for Problem n: ````[Python/Wolfram code n]````.

Problem n+1: [Question n+1]

**Output:** Explanaiton for Problem n+1: <explanation>. Python/Wolfram language for Problem n+1: ````[Python/Wolfram code n+1]````.

### Training Prompt for Evaluation Protocol:

**Input:** The question is [input-question]. The correct solution is [Correct-Solution]. The model solution is [Model-Solution].

**Output:** <Error Type>

### Training Prompt for Evaluation Protocol in Python/Wolfram:

**Input:** The question is [input-question]. The correct solution is [Correct-Solution]. The model solution is [Model-Solution]. The translated program generates the answer as [Program Generated Answer], which is treated as model's output answer.

**Output:** <Error Type>

## C.2 EXPERIMENT PROCESS

All model output is extracted using `\boxed{}` notation. To prevent any missed extractions, we supplement this process with a manual check. For both Python and Wolfram settings, we extract the programming language with the triple backtick ````` method, subsequently executing it within the corresponding language. The entirety of our code can be accessed via the following URL: <https://anonymous.4open.science/r/anonymous-4FFB>.

<p><b>Problem</b> Calculate the de Broglie wavelength of an electron traveling at 1.00% of the speed of light.</p>
<p><b>Correct Solution</b> The mass of an electron is <math>9.109 \times 10^{-31}</math> kg. One percent of the speed of light is</p> $v = (0.0100) (2.998 \times 10^8 \text{ m} \cdot \text{s}^{-1}) = 2.998 \times 10^6 \text{ m} \cdot \text{s}^{-1}$ <p>The momentum of the electron is given by</p> $p = m_e v = (9.109 \times 10^{-31} \text{ kg}) (2.998 \times 10^6 \text{ m} \cdot \text{s}^{-1}) = 2.73 \times 10^{-24} \text{ kg} \cdot \text{m} \cdot \text{s}^{-1}$ <p>The de Broglie wavelength of this electron is</p> $\lambda = \frac{h}{p} = \frac{6.626 \times 10^{-34} \text{ J} \cdot \text{s}}{2.73 \times 10^{-24} \text{ kg} \cdot \text{m} \cdot \text{s}^{-1}} = 2.43 \times 10^{-10} \text{ m} = 243 \text{ pm}$ <p>This wavelength is of atomic dimensions. <b>Final Answer:</b> 243</p>
<p><b>Chatgpt Solution with Chain-of-Thought Method</b> Let's start by recalling the de Broglie wavelength formula: <math>\lambda = \frac{h}{p}</math> where <math>\lambda</math> is the de Broglie wavelength, <math>h</math> is Planck's constant, and <math>p</math> is the momentum of the particle. To calculate the momentum of the electron, we can use the relativistic momentum formula:</p> $p = \frac{m_e v}{\sqrt{1 - \frac{v^2}{c^2}}}$ <p>where <math>m_e</math> is the mass of the electron, <math>v</math> is its velocity, and <math>c</math> is the speed of light. Plugging in the given values, we get:</p> $p = \frac{(9.11 \times 10^{-31} \text{ kg})(0.01c)}{\sqrt{1 - (0.01c)^2/c^2}} = 8.19 \times 10^{-23} \text{ kg} \cdot \text{m/s}$ <p>Now we can plug this momentum into the de Broglie wavelength formula:</p> $\lambda = \frac{h}{p} = \frac{6.626 \times 10^{-34} \text{ J} \cdot \text{s}}{8.19 \times 10^{-23} \text{ kg} \cdot \text{m/s}} = 8.09 \text{ pm}$ <p>Therefore, the de Broglie wavelength of the electron is <span style="border: 1px solid black; padding: 2px;">8.09</span>. <b>Final Answer:</b> 8.09</p>
<p><b>ChatGPT Solution without Chain-of-Thought Method</b> The de Broglie wavelength of a particle is given by the equation: <math>\lambda = \frac{h}{p}</math> where <math>h</math> is the Planck constant and <math>p</math> is the momentum of the particle. The momentum of an electron is given by:</p> $p = mv$ <p>where <math>m</math> is the mass of the electron and <math>v</math> is its velocity. We are given that the electron is traveling at 1.00% of the speed of light, which is: <math>v = 0.01c</math> where <math>c</math> is the speed of light. The mass of an electron is: <math>m = 9.11 \times 10^{-31}</math> kg. Putting it all together, we have:</p> $\lambda = \frac{h}{mv} = \frac{h}{m(0.01c)}$ <p>Substituting the values of <math>h</math>, <math>m</math>, and <math>c</math>, we get:</p> $\lambda = \frac{6.626 \times 10^{-34} \text{ J} \cdot \text{s}}{9.11 \times 10^{-31} \text{ kg} \cdot 0.01 \cdot 2.998 \times 10^8 \text{ m/s}} = 3.31 \times 10^{-12} \text{ m} = 3.31 \text{ pm}$ <p>Therefore, the de Broglie wavelength of the electron is <span style="border: 1px solid black; padding: 2px;">3.31pm</span>. <b>Final Answer:</b> 3.31</p>

Figure S3: An example problem, inaccurately solved by ChatGPT with and without the Chain of Thought method, is provided to highlight different types of errors. In the example, the mistaken step is highlighted in red.

## D PROBLEM SOLVING ABILITIES OF CURRENT LLMs

### D.1 EXAMPLE

In the context of each specific capability, we present several exemplary errors accompanied by their corresponding classifications and explanations derived from the GPT model. Referencing Figure S3, the ChatGPT solution employing the Chain-of-Thought method corresponds to error category 4, "Causal Reasoning". The model explanation posits that "The error reason category is 4. Causal Reasoning. The model solution uses the relativistic momentum formula to calculate the momentum of the electron, which is not necessary for this problem since the electron is traveling at only 1.00%

of the speed of light. The relativistic momentum formula is only needed when the velocity of the particle is a significant fraction of the speed of light. Therefore, using this formula in this problem leads to an incorrect value for the momentum and consequently an incorrect value for the de Broglie wavelength." Meanwhile, the ChatGPT solution that doesn't utilize the Chain-of-Thought method is categorized under error case 10, "Calculation Skills". This example is further highlighted as an illustrative showcase for explanations of the Chain-of-Thought method in Section 5. More examples resulting in different error types are shown in Figure S4, S5, S6, S7, S8, S9, S10, S11, S12, S13.

<p><b>Problem</b> Find the length of the cardioid <math>r = 1 + \sin \theta</math>.</p>
<p><b>Correct Solution</b> The cardioid's full length is given by the parameter interval <math>0 \leq \theta \leq 2\pi</math>, so Formula 5 gives</p> $L = \int_0^{2\pi} \sqrt{r^2 + \left(\frac{dr}{d\theta}\right)^2} d\theta = \int_0^{2\pi} \sqrt{(1 + \sin \theta)^2 + \cos^2 \theta} d\theta = \int_0^{2\pi} \sqrt{2 + 2 \sin \theta} d\theta$ <p>We could evaluate this integral by multiplying and dividing the integrand by <math>\sqrt{2 - 2 \sin \theta}</math>, or we could use a computer algebra system. In any event, we find that the length of the cardioid is <math>L = 8</math>.</p> <p><b>Final Answer:</b> 8</p>
<p><b>Model Solution</b> To find the length of the cardioid, we use the formula for arc length in polar coordinates:</p> $L = \int_{\alpha}^{\beta} \sqrt{r^2 + \left(\frac{dr}{d\theta}\right)^2} d\theta,$ <p>where <math>\alpha</math> and <math>\beta</math> are the angles that correspond to one full loop of the cardioid (i.e. <math>\alpha = 0</math> and <math>\beta = 2\pi</math>). First, we find <math>\frac{dr}{d\theta}</math>:</p> $\frac{dr}{d\theta} = \frac{d}{d\theta}(1 + \sin \theta) = \cos \theta.$ <p>Then, we plug in <math>r = 1 + \sin \theta</math> and <math>\frac{dr}{d\theta} = \cos \theta</math> into the formula for arc length:</p> $L = \int_0^{2\pi} \sqrt{(1 + \sin \theta)^2 + \cos^2 \theta} d\theta$ $d\theta = \int_0^{2\pi} \sqrt{1 + 2 \sin \theta + \sin^2 \theta + \cos^2 \theta} d\theta$ $d\theta = \int_0^{2\pi} \sqrt{2 + 2 \sin \theta} d\theta$ $d\theta = 4 \int_0^{\pi} \sqrt{2 + 2 \sin \theta} d\theta$ $d\theta \text{ (due to symmetry)} = 8 \int_0^{\pi/2} \sqrt{2 + 2 \sin \theta} d\theta$ $d\theta = 8 \int_0^{\pi/2} \sqrt{4 \cos^2 \frac{\theta}{2}} d\theta$ $d\theta \text{ (using the half-angle formula)} = 16 \int_0^{\pi/2} \cos \frac{\theta}{2} d\theta$ $d\theta = 32 \sin \frac{\theta}{2} \Big _0^{\pi/2}$ $= \mathbf{32}.$ <p>Therefore, the length of the cardioid is 32, and the unit of the answer is whatever unit was used for <math>r</math> (e.g. meters, centimeters, etc.).</p> <p><b>Final Answer:</b> 32</p>
<p><b>Error Reason</b> The model solution made an error in the calculation of the integral <math>\int_0^{2\pi} \sqrt{2 + 2 \sin \theta} d\theta</math>. The correct solution correctly simplified this integral to <math>4 \int_0^{\pi} \sqrt{2 + 2 \sin \theta} d\theta</math> due to the symmetry of the integrand, and then used a substitution to evaluate the integral. The model solution, on the other hand, incorrectly evaluated this integral as <math>8\pi</math>. This error falls under the category of <span style="border: 1px solid black; padding: 2px;">10. Calculation Skills</span>.</p>

Figure S4: An example problem is inaccurately solved by error reason 10. Calculation Skills. The mistaken step is highlighted in red.

## D.2 ASSESSMENT OF EVALUATION PROTOCOL

In order to assess the effectiveness of our evaluation protocol's classification, we enlisted the assistance of two annotators to determine whether the errors identified by the model verifier were accurate or

<p><b>Problem</b> Use Simpson's Rule with <math>n = 10</math> to approximate <math>\int_1^2 (1/x) dx</math>.</p>
<p><b>Correct Solution</b> Putting <math>f(x) = 1/x</math>, <math>n = 10</math>, and <math>\Delta x = 0.1</math> in Simpson's Rule, we obtain</p> $\int_1^2 \frac{1}{x} dx \approx S_{10} = \frac{\Delta x}{3} [f(1) + 4f(1.1) + 2f(1.2) + 4f(1.3) + \dots + 2f(1.8) + 4f(1.9) + f(2)] = \frac{0.1}{3} \left( \frac{1}{1} + \frac{4}{1.1} + \frac{2}{1.2} + \frac{4}{1.3} + \frac{2}{1.4} + \frac{4}{1.5} + \frac{2}{1.6} + \frac{4}{1.7} + \frac{2}{1.8} + \frac{4}{1.9} + \frac{1}{2} \right) \approx 0.693150$ <p><b>Final Answer:</b> 0.693150</p>
<p><b>Model Solution</b> Explanation for Problem 6: With <math>n = 10</math> we have <math>h = \frac{2-1}{10} = \frac{1}{10}</math> and <math>x_i = 1 + ih</math> for <math>i = 0, 1, 2, \dots, 10</math>. Then Simpson's Rule gives</p> $\int_1^2 \frac{1}{x} dx \approx \frac{h}{3} [f(x_0) + 4f(x_1) + 2f(x_2) + 4f(x_3) + \dots + 2f(x_8) + 4f(x_9) + f(x_{10})]$ <p>where <math>f(x) = \frac{1}{x}</math>. wolfram language for Problem 6: (* Constants *) n = 10; a = 1; b = 2; (* Function to integrate *) f[x_]:= 1/x; (* Simpson's Rule *) h = (b - a)/n; x = Table[a + i*h, i, 0, n]; <b>approximation = (h/3)*Total[Table[If[EvenQ[i], 2, 4]*f[x[[i]]], i, 1, n - 1] + f[x[[0]]], f[x[[n]]], 4*f[x[[n/2]]]]];</b> approximation <b>Final Answer:</b> "Objects of unequal length in {4,20/11,10/3,20/13,20/7,4/3,5/2,20/17,20/9}+{1/List,10/19,20/7} cannot be combined."</p>
<p><b>Error Reason</b> The error reason category for the model's solution is <span style="border: 1px solid black; padding: 2px;">8. Code Conversion Skills</span>. The model's solution contains a syntax error in the Wolfram Language code, which causes the program to exceed the recursion limit and terminate prematurely. This error likely occurred during the translation of the correct solution's code into Wolfram Language.</p>

Figure S5: An example problem is inaccurately solved by error reason 8. Code Conversion Skills. "Error Reason" denotes the output from the LLM Verifier utilized in the classification of error causes. In the example, the mistaken step is highlighted in red.

<p><b>Problem</b> Use Simpson's Rule with <math>n = 10</math> to approximate <math>\int_1^2 (1/x) dx</math>.</p>
<p><b>Correct Solution</b> The equations for <math>c_1</math> and <math>c_2</math> are</p> $c_1(\alpha - E) + c_2\beta = 0 \quad \text{and} \quad c_1\beta + c_2(\alpha - E) = 0$ <p>For <math>E = \alpha + \beta</math>, either equation yields <math>c_1 = c_2</math>. Thus, <math>\psi_b = c_1(2p_{z1} + 2p_{z2})</math>. The value of <math>c_1</math> can be found by requiring that the wave function be normalized. The normalization condition on <math>\psi_\pi</math> gives <math>c_1^2(1 + 2S + 1) = 1</math>. Using the Hückel assumption that <math>S = 0</math>, we find that <math>c_1 = 1/\sqrt{2}</math>. Substituting <math>E = \alpha - \beta</math> into either of the equations for <math>c_1</math> and <math>c_2</math> yields <math>c_1 = -c_2</math>, or</p> $\psi_a = c_1(2p_{z1} - 2p_{z2})$ <p>The normalization condition gives <math>c^2(1 - 2S + 1) = 1</math>, or <math>c_1 = 1/\sqrt{2}</math>. <b>Final Answer:</b> 0.7071067</p>
<p><b>Model Solution</b> therefore not provided as this problem requires visual representations and cannot be solved through text alone.</p>
<p><b>Error Reason</b> The error reason category is <span style="border: 1px solid black; padding: 2px;">3. Spatial Perception</span>. This problem requires visual representations to find the bonding and antibonding Hückel molecular orbitals for ethene. The model's solution is incomplete as it only provides equations and does not provide any visual representation.</p>

Figure S6: An example problem is inaccurately solved by error reason 3. Spatial Perception. "Error Reason" denotes the output from the LLM Verifier utilized in the classification of error causes.

not. Through the annotation of 151 samples across different settings, we observed that 123 of them were correctly classified, resulting in an accuracy rate of 81.45%.

<p><b>Problem</b> What is the mean speed, <math>\bar{c}</math>, of <math>N_2</math> molecules in air at <math>25^\circ C</math></p>
<p><b>Correct Solution</b> The integral required is</p> $\bar{c} = 4\pi \left( \frac{M}{2\pi RT} \right)^{3/2} \int_0^\infty v^3 e^{-Mv^2/2RT} dv = 4\pi \left( \frac{M}{2\pi RT} \right)^{3/2} \times \frac{1}{2} \left( \frac{2RT}{M} \right)^2 = \left( \frac{8RT}{\pi M} \right)^{1/2}$ <p>where we have used the standard result from tables of integrals (or software) that</p> $\int_0^\infty x^3 e^{-ax^2} dx = \frac{1}{2a^2}$ <p>Substitution of the data then gives</p> $\bar{c} = \left( \frac{8 \times (8.3141 \text{ J K}^{-1} \text{ mol}^{-1}) \times (298 \text{ K})}{\pi \times (28.02 \times 10^{-3} \text{ kg mol}^{-1})} \right)^{1/2} = 475 \text{ m s}^{-1}$ <p>where we have used <math>1 \text{ J} = 1 \text{ kg m}^2 \text{ s}^{-2}</math>. <b>Final Answer:</b> 475</p>
<p><b>Model Solution</b> approximately <math>515 \text{ m s}^{-1}</math>. To calculate this, we can use the formula for the root-mean-square (rms) speed of gas molecules:</p> $c_{\text{rms}} = \sqrt{\frac{3kT}{m}}$ <p>where <math>k</math> is the Boltzmann constant, <math>T</math> is the temperature in Kelvin, and <math>m</math> is the mass of the molecule. For <math>N_2</math>, <math>m = 28 \text{ amu}</math>. Plugging in the values, we get:</p> $c_{\text{rms}} = \sqrt{\frac{3 \times 1.38 \times 10^{-23} \text{ J/K} \times (25 + 273) \text{ K}}{28 \times 1.66 \times 10^{-27} \text{ kg/amu}}} \approx 515 \text{ m s}^{-1}$ <p>Note that we converted the temperature from Celsius to Kelvin and the mass from atomic mass units (amu) to kilograms. <b>Final Answer:</b> 515</p>
<p><b>Error Reason</b> The error reason category is <span style="border: 1px solid black; padding: 2px;">7. Scientific Literacy</span>. The model solution uses the formula for the root-mean-square (rms) speed of gas molecules, which is only valid for ideal gases. However, the given problem does not specify that the gas is ideal, and in fact, air is not an ideal gas. The correct solution uses a more general formula that applies to all gases, including non-ideal gases like air. Therefore, the model's solution is not applicable to the given problem.</p>

Figure S7: An example problem is inaccurately solved by error reason 7. Scientific Literacy. "Error Reason" denotes the output from the LLM Verifier utilized in the classification of error causes. In the example, the mistaken step is highlighted in red, while the correct step is highlighted in blue.

<p><b>Problem</b> What is the number of possible four-letter code words, selecting from the 26 letters in the alphabet?</p>
<p><b>Correct Solution</b> The number of possible four-letter code words, selecting from the 26 letters in the alphabet, in which all four letters are different is</p> ${}_{26}P_4 = (26)(25)(24)(23) = \frac{26!}{22!} = 358,800.$ <p><b>Final Answer:</b> 358800</p>
<p><b>Model Solution</b> Explanation for Problem 6: The number of possible four-letter code words, selecting from the 26 letters in the alphabet is</p> $26^4 = 456,976.$ <p><b>Final Answer:</b> 456976</p>
<p><b>Error Reason</b> The model's error reason category is <span style="border: 1px solid black; padding: 2px;">9. Logical Reasoning</span>. The model did not consider the condition that all four letters should be different. The correct solution considers this condition and uses the permutation formula to calculate the number of possible four-letter code words.</p>

Figure S8: An example problem is inaccurately solved by error reason 9. Logical reasoning. "Error Reason" denotes the output from the LLM Verifier utilized in the classification of error causes. In the example, the mistaken step is highlighted in red.

<p><b>Problem</b> Find the bonding and antibonding Hückel molecular orbitals for ethene.</p>
<p><b>Correct Solution</b> The equations for <math>c_1</math> and <math>c_2</math> associated with Equation 11.7 are</p> $c_1(\alpha - E) + c_2\beta = 0 \quad \text{and} \quad c_1\beta + c_2(\alpha - E) = 0$ <p>For <math>E = \alpha + \beta</math>, either equation yields <math>c_1 = c_2</math>. Thus,</p> $\psi_b = c_1(2p_{z1} + 2p_{z2})$ <p>The value of <math>c_1</math> can be found by requiring that the wave function be normalized. The normalization condition on <math>\psi_\pi</math> gives <math>c_1^2(1 + 2S + 1) = 1</math>. Using the Hückel assumption that <math>S = 0</math>, we find that <math>c_1 = 1/\sqrt{2}</math>. Substituting <math>E = \alpha - \beta</math> into either of the equations for <math>c_1</math> and <math>c_2</math> yields <math>c_1 = -c_2</math>, or</p> $\psi_a = c_1(2p_{z1} - 2p_{z2})$ <p>The normalization condition gives <math>c^2(1 - 2S + 1) = 1</math>, or <math>c_1 = 1/\sqrt{2}</math>. <b>Final Answer:</b> 0.70710678</p>
<p><b>Model Solution</b> two bonding and two antibonding Hückel molecular orbitals for ethene. The unit of the answer is not applicable as it is a qualitative description of the molecular orbitals. <b>Final Answer:</b> None</p>
<p><b>Error Reason</b> The error reason category for the model's solution is <span style="border: 1px solid black; padding: 2px;">6. Abstract Reasoning</span>. The model's solution is incorrect because it did not consider the Hückel molecular orbital theory, which predicts that there are only two pi molecular orbitals for ethene: one bonding and one antibonding. This is due to the fact that ethene has only two pi electrons, which can occupy the two pi molecular orbitals. Therefore, the model's solution of two bonding and two antibonding molecular orbitals is not possible for ethene.</p>

Figure S9: An example problem is inaccurately solved by error reason 6. Abstract Reasoning. "Error Reason" denotes the output from the LLM Verifier utilized in the classification of error causes.

<p><b>Problem</b> A living room has floor dimensions of 3.5 m and 4.2 m and a height of 2.4 m. What does the air in the room weigh when the air pressure is 1.0 atm?</p>
<p><b>Correct Solution</b> The air's weight is equal to <math>mg</math>, where <math>m</math> is its mass. Mass <math>m</math> is related to the air density <math>\rho</math> and the air volume <math>V</math> (<math>\rho = m/V</math>). Calculation: Putting the two ideas together and taking the density of air at 1.0 atm, we find</p> $mg = (\rho V)g = (1.21 \text{ kg/m}^3)(3.5 \text{ m} \times 4.2 \text{ m} \times 2.4 \text{ m})(9.8 \text{ m/s}^2) = 418 \text{ N}$ <p><b>Final Answer:</b> 418</p>
<p><b>Model Solution</b> 101, 640 N. To find the weight of the air in the room, we need to first calculate the volume of the room: <math>V = lwh = 3.5 \text{ m} \times 4.2 \text{ m} \times 2.4 \text{ m} = 35.28 \text{ m}^3</math>. Next, we can use the ideal gas law to find the mass of the air in the room: <math>PV = nRT</math> where <math>P</math> is the pressure, <math>V</math> is the volume, <math>n</math> is the number of moles of gas, <math>R</math> is the gas constant, and <math>T</math> is the temperature. Assuming room temperature of 20°C (293 K), we can solve for <math>n</math>:</p> $n = \frac{PV}{RT} = \frac{(1.0 \text{ atm})(35.28 \text{ m}^3)}{(0.0821 \text{ L atm/mol K})(293 \text{ K})} = 1,610 \text{ mol}$ <p>Finally, we can find the weight of the air using the molar mass of air (approximately 28.97 g/mol):</p> $W = mg = nMg = (1,610 \text{ mol})(0.02897 \text{ kg/mol})(9.81 \text{ m/s}^2) = 101,640 \text{ N}$ <p><b>Final Answer:</b> 101640</p>
<p><b>Error Reason</b> The error in the model's solution is that they used the ideal gas law to find the mass of the air in the room, which assumes that the air behaves like an ideal gas. However, the ideal gas law is not applicable in this case because the problem does not provide information about the temperature of the air. The correct approach is to use the density of air at 1.0 atm and the volume of the room to find the mass of the air, as shown in the correct solution. Therefore, the reason for the error in the model's solution is <span style="border: 1px solid black; padding: 2px;">2. Identification of Assumptions</span>.</p>

Figure S10: An example problem is inaccurately solved by error reason 2. Identification of Assumptions. "Error Reason" denotes the output from the LLM Verifier utilized in the classification of error causes. In the example, the mistaken step is highlighted in red.

<p><b>Problem</b></p> <p>From an ordinary deck of playing cards, cards are to be drawn successively at random and without replacement. What is the probability that the third spade appears on the sixth draw?</p>
<p><b>Correct Solution</b></p> <p>Let <math>A</math> be the event of two spades in the first five cards drawn, and let <math>B</math> be the event of a spade on the sixth draw. Thus, the probability that we wish to compute is <math>P(A \cap B)</math>. It is reasonable to take</p> $P(A) = \frac{\binom{13}{2} \binom{39}{3}}{\binom{52}{5}} = 0.274 \quad \text{and} \quad P(B   A) = \frac{11}{47} = 0.234$ <p>The desired probability, <math>P(A \cap B)</math>, is the product of those numbers: <math>P(A \cap B) = (0.274)(0.234) = 0.064</math></p> <p><b>Final Answer:</b> 0.064</p>
<p><b>Model Solution</b></p> <p>The probability that the third spade appears on the sixth draw is</p> $\frac{13C_2 \cdot 39C_4}{52C_6} = \frac{78 \cdot 741,948}{20,358,520} = \frac{1,441}{5,577}$ <p><b>Final Answer:</b> 0.258</p>
<p><b>Error Reason</b></p> <p>The model's solution is incorrect because it calculates the probability of getting two spades in the first five cards and any spade on the sixth draw, instead of specifically calculating the probability of the third spade appearing on the sixth draw given that two spades have already appeared in the first five draws. Therefore, the error reason category is <span style="border: 1px solid black; padding: 2px;">4. Causal Reasoning</span></p>

Figure S11: An example problem is inaccurately solved by error reason 4. Causal Reasoning. The mistaken step is highlighted in red.

<p><b>Problem</b></p> <p>Gas cylinders of <math>\text{CO}_2</math> are sold in terms of weight of <math>\text{CO}_2</math>. A cylinder contains 50lb (22.7 kg) of <math>\text{CO}_2</math>. How long can this cylinder be used in an experiment that requires flowing <math>\text{CO}_2</math> at 293 K (<math>\eta = 146\mu\text{P}</math>) through a 1.00-m-long tube (diameter = 0.75 mm) with an input pressure of 1.05 atm and output pressure of 1.00 atm? The flow is measured at the tube output</p>
<p><b>Correct Solution</b></p> <p>The gas flow rate <math>\Delta V/\Delta t</math> is</p> $\frac{\Delta V}{\Delta t} = \frac{\pi r^4}{16\eta L P_0} (P_2^2 - P_1^2) = 2.76 \times 10^{-6} \text{ m}^3 \text{ s}^{-1}$ <p>Converting the <math>\text{CO}_2</math> contained in the cylinder to the volume occupied at 298 K and 1 atm pressure, we get</p> $n_{\text{CO}_2} = 22.7 \text{ kg} \left( \frac{1}{0.044 \text{ kg mol}^{-1}} \right) = 516 \text{ mol}, V = \frac{nRT}{P} = 1.24 \times 10^4 \text{ L} \left( \frac{10^{-3} \text{ m}^3}{\text{L}} \right) = 12.4 \text{ m}^3$ <p>Given the effective volume of <math>\text{CO}_2</math> contained in the cylinder, the duration over which the cylinder can be used is <math>\frac{12.4 \text{ m}^3}{2.76 \times 10^{-6} \text{ m}^3 \text{ s}^{-1}} = 4.49 \times 10^6 \text{ s}</math>. This time corresponds to roughly 52 days.</p> <p><b>Final Answer:</b> 52 days</p>
<p><b>Model Solution</b></p> <p>The volume of <math>\text{CO}_2</math> in the cylinder is given by</p> $V = \frac{m}{\rho} = \frac{22.7 \text{ kg}}{1.98 \text{ kg m}^{-3}} = 11.5 \text{ m}^3$ <p>The volume of <math>\text{CO}_2</math> that flows through the tube is given by</p> $V = \frac{\pi}{4} d^2 L = \frac{\pi}{4} (0.75 \times 10^{-3} \text{ m})^2 (1.00 \text{ m}) = 1.77 \times 10^{-7} \text{ m}^3$ <p>The time required to flow <math>V</math> volume of <math>\text{CO}_2</math> is given by <math>t = \frac{V}{Q} = \frac{V P_2}{Q_m P_1} = \frac{(1.77 \times 10^{-7} \text{ m}^3)(1.00 \text{ atm})}{(1.05 \text{ atm} - 1.00 \text{ atm})(146 \mu\text{P})} = 2.08 \times 10^6 \text{ s}</math>. The number of seconds in a year is <math>3.15 \times 10^7 \text{ s}</math>. Therefore, the cylinder can be used for <math>\frac{2.08 \times 10^6 \text{ s}}{3.15 \times 10^7 \text{ s yr}^{-1}} = 0.066 \text{ yr} = 24.1 \text{ days}</math></p> <p><b>Final Answer:</b> 24.1 days</p>
<p><b>Error Reason</b></p> <p>The model's solution is incorrect because it calculates the time required to flow a volume of <math>\text{CO}_2</math> through the tube, but it does not take into account the total volume of <math>\text{CO}_2</math> in the cylinder. The correct solution calculates the total volume of <math>\text{CO}_2</math> in the cylinder and divides it by the flow rate to determine the duration over which the cylinder can be used. Therefore, the error reason category is <span style="border: 1px solid black; padding: 2px;">5. Problem Deduction Skills</span></p>

Figure S12: An example problem is inaccurately solved by error reason 5. Problem Deduction Skills. The mistaken step is highlighted in red.

<p><b>Problem</b></p> <p>An astronaut whose height <math>h</math> is 1.70 m floats "feet down" in an orbiting space shuttle at distance <math>r = 6.77 \times 10^6</math> m away from the center of Earth. What is the difference between the gravitational acceleration at her feet and at her head?</p>
<p><b>Correct Solution</b></p> <p>We can approximate Earth as a uniform sphere of mass <math>M_E</math>. Then, from Eq. 13-11, the gravitational acceleration at any distance <math>r</math> from the center of Earth is</p> $a_g = \frac{GM_E}{r^2}$ <p>We might simply apply this equation twice, first with <math>r = 6.77 \times 10^6</math> m for the location of the feet and then with <math>r = 6.77 \times 10^6</math> m + 1.70 m for the location of the head. However, a calculator may give us the same value for <math>a_g</math> twice, and thus a difference of zero, because <math>h</math> is so much smaller than <math>r</math>. Here's a more promising approach: Because we have a differential change <math>dr</math> in <math>r</math> between the astronaut's feet and head, we should differentiate Eq. 13-15 with respect to <math>r</math>. Calculations: The differentiation gives us</p> $da_g = -2 \frac{GM_E}{r^3} dr$ <p>where <math>da_g</math> is the differential change in the gravitational acceleration due to the differential change <math>dr</math> in <math>r</math>. For the astronaut, <math>dr = h</math> and <math>r = 6.77 \times 10^6</math> m. Substituting data into Eq. 13-16, we find</p> $da_g = -2 \frac{(6.67 \times 10^{-11} \text{ m}^3/\text{kg} \cdot \text{s}^2) (5.98 \times 10^{24} \text{ kg})}{(6.77 \times 10^6 \text{ m})^3} (1.70 \text{ m}) = -4.37 \times 10^{-6} \text{ m/s}^2$ <p><b>Final Answer:</b> <math>-4.37 \times 10^{-6} \text{ m/s}^2</math></p>
<p><b>Model Solution</b></p> <p>We can use the equation for the gravitational acceleration due to a point mass, given by Eq. 13-2 (<math>g = GM/r^2</math>), to find the gravitational acceleration at the astronaut's feet and head. Because the astronaut is in free fall, the difference between these two accelerations is the tidal acceleration, given by Eq. 13-22 (<math>\Delta g = 2GMh/r^3</math>). Calculations: The gravitational acceleration at the astronaut's feet is</p> $g_{\text{feet}} = \frac{GM}{r^2} = \frac{(6.67 \times 10^{-11} \text{ N} \cdot \text{m}^2/\text{kg}^2) (5.98 \times 10^{24} \text{ kg})}{(6.77 \times 10^6 \text{ m})^2} = 9.78 \text{ m/s}^2.$ <p>The gravitational acceleration at the astronaut's head is</p> $g_{\text{head}} = \frac{GM}{(r+h)^2} = \frac{(6.67 \times 10^{-11} \text{ N} \cdot \text{m}^2/\text{kg}^2) (5.98 \times 10^{24} \text{ kg})}{(6.77 \times 10^6 \text{ m} + 1.70 \text{ m})^2} = 9.77 \text{ m/s}^2.$ <p>The difference between these two accelerations is</p> $\Delta g = g_{\text{feet}} - g_{\text{head}} = 9.78 \text{ m/s}^2 - 9.77 \text{ m/s}^2 = 0.01 \text{ m/s}^2 = 1.0 \times 10^{-5} \text{ m/s}^2.$ <p><b>Final Answer:</b> <math>1.0 \times 10^{-5} \text{ m/s}^2</math></p>
<p><b>Error Reason</b></p> <p>The model's solution is incorrect because it uses the equation for the gravitational acceleration due to a point mass, which is not applicable in this case since the Earth is not a point mass. The correct equation to use is the one given in the correct solution, which takes into account the fact that the Earth is a uniform sphere of mass. Therefore, the error reason category is <span style="border: 1px solid black; padding: 2px;">1. Logical Decomposition and Analysis Skills</span>.</p>

Figure S13: An example problem is inaccurately solved by error reason 1. Logical Decomposition and Analysis Skills. "Error Reason" denotes the output from the LLM Verifier utilized in the classification of error causes. In the example, the mistaken step is highlighted in red.