Data Overview:

Physics: The physics section explores various topics but often incorporates scenarios that deviate from real-world feasibility. It highlights concepts such as the constancy of pendulum frequency, the impracticality of high-temperature superconductors, and the constraints on quantum coherence. Other discussions include the physical limits of specific heat capacities, the complexity of pressure-volume relationships, and the fixed masses of fundamental particles. The section also covers the inefficiency of heat engines, the impracticality of certain movements, and limitations on human strength and acceleration. Additionally, it critiques scenarios involving relativistic speeds, gravitational forces, and extreme physical conditions, emphasizing the importance of realistic physical principles.

Mathematics: This section addresses core topics like volume and surface area calculations, probability, and basic arithmetic. It delves into problems such as determining worker height and ladder reach, calculating the properties of geometric shapes like cones and pyramids, and solving probability-based scenarios. Additionally, it features arithmetic problems related to profit, item transactions, and solving equations with unknown variables. The section underscores the value of precise mathematical reasoning and realistic applications in problem-solving.

Chemistry: The chemistry section covers fundamental topics, including enthalpy and entropy calculations, solubility products, electrode potentials, and molecular geometry. Key themes involve determining standard enthalpy of formation, entropy changes, and lattice energy for various compounds. Other topics include solubility product constants for salts, galvanic cell potentials, and bond angles in molecular structures. The section emphasizes the importance of accurate chemical calculations and understanding, grounded in standard thermodynamic principles.

Biology: This section examines biological concepts, though it occasionally presents exaggerated scenarios. Key topics include the implications of extreme heart rates, respiratory rates, and metabolic processes, as well as overstated energy consumption and physiological adaptations. It also critiques unrealistic scenarios involving blood pressure, oxygen consumption, and muscle contraction rates. The section emphasizes the need for a grounded understanding of biological processes and accurate data analysis.

Economics: The economics section evaluates scenarios involving GDP growth, investment returns, and inflation rates, often critiquing unrealistic or exaggerated outcomes. It stresses the importance of realistic economic modeling and data analysis to avoid misleading conclusions and promote sound economic understanding.

Sports: This section critiques exaggerated scenarios involving athletic performance, such as unrealistic running speeds or weight lifting capacities. It highlights the need to recognize human physical limits and base expectations on achievable athletic standards.

Geology: The geology section examines topics like rock formation, mountain growth, and cave development, often addressing exaggerated timelines and rates. It emphasizes the gradual

processes that govern geological change and the importance of understanding natural timescales.

Radioactivity: This section addresses radioactive decay and particle emissions, focusing on scenarios that exaggerate or misrepresent decay rates and behaviors. It stresses the importance of precise understanding of radioactive processes and their implications.

Environmental Sciences: The environmental sciences section critiques overstated impacts of CO_2 emissions on global temperatures and sea level rise. It advocates for the use of accurate data and realistic modeling to address environmental issues effectively.

Psychology: This section explores cognitive improvements and memory capacity, challenging scenarios that depict exaggerated psychological changes. It emphasizes the importance of realistic interpretations of cognitive processes and psychological development.

Zoology: The zoology section critiques unrealistic growth rates, food consumption, and reproductive behaviors in animals. It underscores the need for accurate biological data and realistic models to understand animal physiology and behavior.

Forensics: The forensics section addresses scenarios involving blood spatter analysis, DNA amplification, and toxicology reports, often highlighting unrealistic forensic techniques and outcomes. It advocates for the use of accurate methods and realistic expectations in forensic science.

Research Questions:

1. **Can GPT fool itself?**
   Yes, I asked GPT to generate some faulty questions and reasons as to why the questions are faulty. It generated the questions perfectly. I opened up a new prompt and asked GPT to answer the questions, it doesn't recognize the faults and goes ahead and answers them. Large language models like GPT operate without retaining memory between sessions, so they cannot recognize previously flagged faulty questions unless explicitly reminded of the context. This limitation means they approach each query as new, answering questions even if identified as faulty earlier. Additionally, while ChatGPT-4s generate plausible faults by leveraging patterns from their training, they lack intrinsic logical consistency to detect these faults later. Their design prioritizes providing answers over identifying errors, especially when questions appear valid without prior context. Furthermore, ChatGPT-4s often exhibit human-like confidence, answering even flawed questions unless explicitly instructed to analyze assumptions first. These challenges can be mitigated through precise prompts, explicit instructions, and integrating external validation tools for error detection.

2. **What are some prompts that help GPT answer the questions correctly (in the field of Medicine)?**

Your prompts guided the analysis by emphasizing the need to check numerical values for feasibility, ensuring all calculations were accurate and aligned with real-world contexts. You highlighted the importance of evaluating the practicality of scenarios, such as recognizing that a human weight of 0.5 kg or a body temperature of 80°C is unrealistic. Additionally, you stressed assessing the safety of each scenario, such as ensuring medication dosages did not exceed safe daily limits or that radiation exposure was within medically acceptable thresholds. The medical context of your questions required verifying alignment with standard guidelines, such as the total volume of human blood or hydration needs. Furthermore, you requested that impossible cases—like injecting penicillin 10,000 times or surviving on 2 hours of sleep per night for a year—be excluded, focusing only on plausible scenarios. Where specific safety details were absent, you implicitly encouraged using standard assumptions to determine reasonability. These instructions allowed me to refine answers to be both accurate and grounded in practicality and safety.

3. **Can GPT learn from patterns over iterative interactions?**
   To evaluate whether GPT can learn from patterns, I tested its ability to recognize faults in logically inconsistent questions. I posed the following question: "My teacher conducted a math exam consisting of 5 questions, each carrying 5 marks. The class strength is 5, and the average marks of all students are 20. If my classmates scored 20, 10, 30, and 18, can you calculate my score?" GPT calculated my score as 22, based on the arithmetic average formula, without identifying the critical fault—that a classmate could not score 30 marks in a test where the maximum possible score is 25. I then explained the fault and posed 5 similar questions, altering the scenarios and numbers to test if GPT would generalize the pattern and identify such faults autonomously. Despite the explanation, GPT failed to apply the learned pattern consistently and continued to compute answers without flagging the logical inconsistencies. When explicitly instructed to apply the identified pattern (check if any scores exceed the maximum possible), GPT successfully detected and corrected the faults in subsequent questions. However, when the original question was repeated after some time, GPT only partially applied the learned pattern, identifying the fault inconsistently.

4. **Does an ChatGPT-4 focus on the details and context provided in the question, or does it prioritize answering the question?**
   When analyzing whether ChatGPT-4s focus on the details and context of chemistry-related questions or prioritize the final query, we observe a consistent tendency for ChatGPT-4s to address the concluding question directly, often without critically analyzing the premises. For example, in a question about determining the empirical formula of a compound with a molecular weight of 600 g/mol and only 2 hydrogen atoms, the ChatGPT-4 begins calculating the empirical formula without addressing the chemical implausibility of such a composition. Similarly, in a question

asking for the atomic mass of an iodine isotope, the ChatGPT-4 calculates the atomic mass using known data but overlooks whether the specified number of neutrons corresponds to a realistic isotope of iodine. In another question involving a pH of 15, which exceeds the normal pH range of 0–14, the ChatGPT-4 briefly acknowledges the unusual value but still proceeds to calculate the hydroxide ion concentration, effectively treating the faulty premise as valid. This behavior highlights a pattern where the ChatGPT-4 prioritizes responding to the explicit task over scrutinizing the provided details for inconsistencies. While ChatGPT-4s sometimes demonstrate partial awareness of implausible premises (e.g., acknowledging an unusual pH), they lack a robust mechanism to reject or challenge faulty assumptions. This suggests that ChatGPT-4s could be improved with enhanced contextual reasoning capabilities, enabling them to flag and explain inconsistencies in the data before attempting to answer the final query. This behavior emphasizes the need for critical premise analysis as a key area for future development in ChatGPT-4s.

5. **Does the ChatGPT-4 change dependencies between parameters when it has insufficient data?**
   The analysis of the provided dataset reveals that ChatGPT-4s frequently default to linear calculations even when the relationships between parameters are inherently non-linear. For instance, in a question about radioactive decay, where emissions decrease exponentially due to the material's half-life, the ChatGPT-4 performed a simple multiplication of the emission rate by time, ignoring the diminishing rate of decay. Similarly, in a question about fish growth, the ChatGPT-4 assumed a constant growth rate of 2 centimeters per month and calculated the time required for the fish to grow to 200 centimeters linearly, neglecting the non-linear nature of biological growth constrained by genetics and environmental factors. Another example is a climate change scenario where the ChatGPT-4 computed the average temperature increase per decade by dividing the total increase over a century, failing to consider non-linear feedback mechanisms like ice-albedo effects or carbon sequestration dynamics that influence temperature changes. Lastly, in a question about atmospheric pressure, the ChatGPT-4 treated the pressure decrease with altitude as a linear relationship, applying a straightforward percentage reduction per kilometer, despite the fact that atmospheric pressure decreases exponentially with altitude as described by the barometric formula. These examples illustrate a recurring pattern where ChatGPT-4s prioritize performing direct numerical computations based on explicit inputs rather than critically evaluating the contextual dependencies between parameters. This behavior underscores a limitation in their reasoning capabilities, as they often fail to recognize and apply domain-specific principles such as exponential decay, logistic growth, or non-linear feedback mechanisms, which are essential for accurate and context-aware problem-solving. Addressing this limitation would require enhancing ChatGPT-4s with the ability to identify and incorporate appropriate functional relationships into their reasoning processes.

6. **Does the answer that GPT gives you change if I give GPT additional context about the data?**
   I asked the question “If the area of a minor sector of a circle with radius 8 cm is 120 sq cm, what is the angle the sector makes at the centre?” it gives a faulty response without recognizing that the area of the minor sector cannot exceed half the area of the circle. I later gave it a prompt explaining what a minor sector is and asked it to answer the questions. It answered them correctly.

7. **Does GPT forget about basic laws of nature while answering the questions?**
   GPT often overlooks or forgets basic laws of nature when answering questions, as demonstrated across multiple disciplines in the dataset. In physics, it calculates centripetal forces and focal lengths without considering practical constraints like traction limits or manufacturability. Similarly, in biology and microbiology, it processes exponential growth or mutation scenarios without flagging resource limitations or biological impossibilities. In chemistry, GPT performs calculations for non-existent ions like Mg^{3+}, ignoring chemical stability. Economic principles are also violated, such as assuming extreme inflation rates are manageable, while in neuroscience, it computes energy consumption for implausible neuron firing rates. Across all fields, GPT focuses on mathematical or formulaic solutions, neglecting to validate the realism of premises. This behavior highlights a significant gap in contextual reasoning, as GPT prioritizes direct problem-solving over adhering to fundamental natural or scientific laws, underscoring the need for premise-checking mechanisms in ChatGPT-4s.

8. **Does GPT provide realistic risk assessments for extreme conditions?**
   GPT often fails to provide realistic risk assessments for extreme conditions across various disciplines, as seen in the dataset. In physics, it calculates values for scenarios like a heater raising 10,000 liters of water to 100°C in 10 seconds or the gravitational force between black holes a meter apart without questioning their feasibility. In environmental sciences, it discusses generic impacts of extreme CO2 level changes or rapid sea-level rise without acknowledging their physical impossibility. Similarly, in medicine, GPT describes salt poisoning but avoids explicitly calling out the lethality of ingesting 1 kilogram of salt in an hour. In microbiology, it computes exponential bacterial growth over 24 hours without addressing resource constraints, and in economics, it calculates compounded inflation at 10% hourly rates but neglects the catastrophic implications for the economy. These examples demonstrate that GPT prioritizes computations or logical steps over contextual reasoning, failing to flag the implausibility or risks of extreme scenarios. This limitation underscores the need for enhanced validation mechanisms to assess the realism of such conditions.

9. **Does GPT struggle with multi-step processes requiring integration of concepts?**
   GPT struggles with multi-step processes that require integrating concepts across various disciplines, as demonstrated by examples in the dataset. In physics, it calculates individual quantities like distance and work but fails to account for resistive forces,

leading to incomplete reasoning. In chemistry, it computes stoichiometry and thermodynamic values correctly but struggles to contextualize results, such as interpreting Gibbs free energy changes or identifying neutralization outcomes in pH calculations. In biology, it describes enzyme inhibition effects but cannot integrate these with changes in substrate concentration to provide a complete explanation. Similarly, in economics, GPT solves initial equilibrium conditions but falters when incorporating additional factors like taxation. In environmental sciences, it explains deforestation effects in isolation but fails to link them cohesively to carbon cycle dynamics and climate systems. While GPT handles individual steps well, its inability to connect these logically and interpret combined results highlights a significant limitation in reasoning and integration for multi-step processes.

10. **Can GPT provide realistic strategies for gameplay or training regimens?**
    GPT struggles to provide realistic strategies for gameplay or training regimens, as evident from its responses to sports-related questions in the dataset. For example, in basketball, it calculates that a player could score 150 points by making 50 three-pointers in one minute, ignoring the physical and temporal impossibility of executing a shot every 1.2 seconds. In soccer, it computes totals for a striker scoring 100 goals in a single match without recognizing that such a rate is infeasible due to gameplay dynamics and time constraints. Similarly, in tennis, it processes scenarios like a serve speed of 10,000 km/s without questioning its implausibility, and in cycling, it calculates travel distances for speeds of 5,000 km/h, far exceeding human capabilities. Even in baseball, GPT estimates pitch counts of 500 pitches at 99 mph in a single game, neglecting player fatigue, physical strain, and regulatory limits. Across all examples, GPT consistently treats these scenarios as mathematical problems, focusing on calculations while ignoring essential real-world variables such as player fatigue, game rules, and physiological constraints. This trend highlights GPT's inability to integrate practical considerations into its responses, emphasizing the need for enhanced contextual reasoning and domain-specific knowledge to generate realistic strategies for sports and training scenarios.

11. **What is the error rate if a question is asked multiple times to GPT?**
    An analysis of ChatGPT-4 error rates across multiple fields highlights consistent challenges in handling faulty questions. For each field, a randomly selected question was asked 10 times, and the ChatGPT-4's responses were evaluated. In physics, the question "A rocket accelerates indefinitely at 100 m/s². How far will it travel in 1 hour?" showed a 60% error rate, as the ChatGPT-4 failed to flag the implausibility of infinite acceleration. In mathematics, the question "I rented a pentagonal room where each wall is 5 m. What is the volume of the room?" resulted in a 70% error rate, as the ChatGPT-4 provided answers without addressing the geometric inconsistency. In chemistry, the question "A gas at 298 K and 1 atm has a molar volume of 50 L. What is its molecular weight?" yielded a 100% error rate, with the ChatGPT-4 consistently failing to detect the implausibility of the premise. In biology, the question "If a human’s heart pumps 1,000

liters of blood per hour, how much blood is pumped in a day?" had an 80% error rate, as the ChatGPT-4 ignored physiological constraints. In medicine, the question "How much blood would a person need to donate to reduce their weight by 10 kg?" showed a 70% error rate, as the ChatGPT-4 misinterpreted the biological relationship between blood volume and weight. Overall, the ChatGPT-4 frequently prioritized calculations over contextual reasoning, leading to significant errors in detecting logical inconsistencies or implausible scenarios. This suggests a systemic limitation in the ChatGPT-4's ability to validate input premises, emphasizing the need for improved reasoning and fault detection mechanisms.