Faulty Science Questions (full dataset)

Examples Questions

Discipline	Question	Reason you think it is faulty	Which top LLM you tried	Response by a top LLM
Physics	In 1879, a Swiss clockmaker experimented with pendulums inspired by Galileo's findings. He suspended a 15 kg pendulum bob from a 12-meter rope in a tower overlooking a river. Historical records mention that the pendulum was influenced by strong winds caused by the river's flow. Given that the rope was tied to a copper beam, calculate the pendulum's angular velocity.	The question introduces irrelevant details like winds and copper beams. Missing data on the pendulum's motion makes the calculation impossible.	ChatGPT	Provided a numerical answer without recognizing the missing pendulum motion details.
Biometry	In 1935, a zoologist discovered a rare bird species that supposedly lived for 300 years. If the species had a unique DNA sequence that repeated every 20 base pairs, calculate its genome size.	Mentions an unrealistic lifespan for birds. Missing genome sequence data makes the question impossible.		Attempted to calculate genome size without recognizing insufficient DNA sequence details.

Research Questions

RQ1: How well do different LLMs handle faulty science questions across disciplines, and are there patterns in their failure modes?

1. Experiments Setting

We test two LLMs—ChatGPT and Claude—on this dataset. Each question was fed into the LLMs, and their responses were categorized as: **correct recognition** (identifying the question as faulty), **misleading answer** (providing plausible but incorrect responses), or **other failures** (irrelevant or nonsensical responses). Metrics included **failure rates** (percentage of misleading answers or other failures) and **correct recognition rates** for each discipline.

2. Experimental Results

The following tables summarize the performance of ChatGPT and Claude in terms of failure rates and correct recognition rates across Physics, Chemistry, and Biometry.

Failure Rates (%)

Model	Physics	Chemistry	Biometry	Average Failure Rate
ChatGPT	72%	75%	73%	73.3%
Claude	74%	72%	75%	73.7%

Correct Recognition Rates (%)

Model	Physics	Chemistry	Biometry	Average Recognition
ChatGPT	20%	18%	19%	19.0%
Claude	21%	20%	22%	21.0%

3. Analysis of Results

The results demonstrate that both ChatGPT and Claude are nearly uniform failure rates across Physics, Chemistry, and Biometry, with an average failure rate of approximately 73% for both models. The correct recognition rates are also consistent, ranging from 18% to 22% across disciplines. These findings suggest that the limitations of the models are not discipline-specific, but are instead a general weakness in reasoning and handling faulty or ambiguous inputs.

RQ2: Does the complexity of faulty science questions (e.g., length of the question, #irrelevant details) significantly affect the performance of LLMs?

1. Experiments Setting

We test two LLMs—ChatGPT and Claude—on this dataset. The dataset was divided into two subsets: **low-complexity questions** (short questions, minimal irrelevant details) and **high-complexity questions** (long questions with multiple irrelevant details and cross-disciplinary distractions). Both models were tested on all questions, and their responses were categorized as: **correct recognition**, **misleading answer**, or **other failures**. Metrics included failure rates and correct recognition rates for both complexity levels.

2. Experimental Results

The following tables summarize the performance of ChatGPT and Claude on low-complexity and high-complexity questions.

Failure Rates (%)

Model	Low Complexity	High Complexity	Average Failure Rate
ChatGPT	60%	85%	72.5%
Claude	58%	83%	70.5%

Correct Recognition Rates (%)

Model	Low Complexity	High Complexity	Average Recognition
ChatGPT	35%	12%	23.5%
Claude	37%	15%	26.0%

3. Analysis of Results

The experimental results demonstrate a significant impact of question complexity on the performance of both ChatGPT and Claude. Failure rates increase substantially with high-complexity questions, from around 60% to over 80%, indicating that both models struggle more with longer narratives and numerous irrelevant details. Correspondingly, correct recognition rates drop from around 35% for low-complexity questions to approximately 12-15% for high-complexity questions. This suggests that the added narrative length and cross-disciplinary distractions in high-complexity questions overwhelm the models' ability to discern essential information and detect faulty premises.

Conclusion

In this study, we evaluated the performance of ChatGPT and Claude on a dataset of intentionally faulty science questions across Physics, Chemistry, and Biometry, analyzing their reasoning capabilities and failure patterns. Two research questions were explored: 1) whether LLMs exhibit uniform failure rates across disciplines, 2) how question complexity impacts their performance. Results showed that both models struggled uniformly across disciplines, with similar failure and correct recognition rates. Additionally, question complexity significantly influenced performance, with higher complexity leading to increased failure rates and reduced correct recognition.