CSE 584 – Final Project

- **D.Lakshmi Sivani**
- **PSU ID: 906922347**

- **Basic Introduction:**

  Advanced large language models have completely changed problem-solving capabilities in advanced mathematics, science, and general knowledge. However, these systems still are not resistant to subtle errors: they tend to fall for misleading or flawed questions that require reasoning beyond the superficial meaning of a text. Establishing these weaknesses is important for the robustness of LLMs, especially in high-stakes applications like education, scientific research, and decision-making. This work is a follow-up to previous efforts like FaultyMath, which researched ways to create faulty math problems in order to challenge and assess the reasoning of LLMs. Drawing inspiration from the methods in FaultyMath, we expanded that approach to construct a varied dataset of faulty science questions. By harnessing key datasets such as QASPER and SciQ, we prepared a custom dataset for testing the robustness of state-of-the-art LLMs in terms of sound reasoning when faced with deliberately misleading science problems.

  Our objective is twofold: (1) to create a dataset of faulty science questions across diverse categories, ensuring coverage of various scientific domains and problem structures, and (2) to design and conduct experiments to evaluate how well-leading LLMs, such as Gemini 1.5 Flash, GPT-4, and others, can detect and address these faulty problems. This paper describes the curation process for the dataset, the experimental framework, and preliminary results; it forms a starting point toward deeper explorations of LLMs' limitations in reasoning.

- **Dataset Description:**

  The "**Faulty_Q dataset"** is a curated collection of logically flawed and ambiguous questions designed to test the reasoning capabilities of advanced models. The dataset focuses on diverse fault types across multiple domains, ensuring robust evaluation and insights into model limitations.

  This dataset is comprised of 400 faulty questions curated from above-mentioned datasets in different disciples like –

  1) Mathematics - faulty problems related to ambiguous arithmetic scenarios (e.g., invalid operations or setups), logical inconsistencies in ratios and proportions, faulty sequences and series, paradoxes in algebraic expressions, and errors in problem setup for geometry or ratios.

  2) Physics - faulty problems related to Miscalculations in thermal and linear expansion, errors in mechanics (e.g., variable acceleration or velocity problems),

flaws in thermodynamic assumptions (e.g., pressure, volume, and temperature), misinterpretations in gear ratios and rotational mechanics, faulty applications of gravitational and resistance calculations.

3) Chemistry – problems introducing false assumptions in molecular structures (e.g., impossible bond angles), misinterpretation of thermodynamic properties (e.g., entropy or heat transfer), errors in gas laws and volume/temperature relationships, incorrect molecular and chemical bonding theories and Unrealistic physical transformations in matter.

4) Biology – questions related to impossible blood pressure or heart rate readings, erroneous metabolic or circulatory assumptions, Misinterpretation of human body limits (e.g., zero cardiac output survival), Ambiguities in oxygen absorption and blood flow, errors in uniform body system pressure.

5) Astronomy – problems related to errors in celestial mechanics (e.g., impossible planetary orbits), unrealistic expansion, or energy scenarios in universal physics, Misinterpretations of black hole or stellar dynamics, Ambiguities in light and sound propagation in space, and Unrealistic observational setups (e.g., seeing the Sun at night).

6) Psychology – questions related to Unrealistic emotional or mental states (e.g., infinite emotions), faulty memory retention, anxiety reduction models, errors in attention span or stress decay assumptions, ambiguities in habit formation, and dopamine level predictions, exaggerated scenarios for therapy or phobia treatments.

7) Environmental Science – questions related to Unrealistic ecosystem transformations (e.g., rapid desert-to-forest changes), errors in atmospheric chemistry (e.g., oxygen molecule creation rates), ambiguities in global warming or carbon absorption calculations, faulty assumptions in water cycle or energy generation systems and exaggerated ecological phenomena.

8) Geology – problems related to faulty tectonic and volcanic transformations (e.g., extreme plate movements), Unrealistic crystal and fossil formation rates, errors in sedimentary or rock density assumptions, ambiguities in P-wave travel or geological epoch transformations, exaggerated mineral composition generation.

9) Statistics – problems related to misapplication of probabilities and statistical independence, faulty assumptions in normal distributions and confidence intervals, errors in correlation coefficients and regression models, ambiguities in p-value calculations and chi-square tests, misinterpretation of survey data or sampling errors.

10) Economics – problems related to miscalculations in inflation, GDP, or interest rates, faulty employment or labor productivity assumptions, errors in market elasticity and multiplier effects, ambiguities in economic growth predictions and Unrealistic tariff or revenue models.

11) Meteorology – problems to misinterpretations of atmospheric pressure and precipitation rates, errors in hurricane intensity and wind-speed calculations, faulty temperature-relative humidity relationships, ambiguities in solar radiation or snow accumulation models, and Unrealistic weather system dynamics.

12) Miscellaneous/General Science – problems related to faulty interpretations of astronomical and geological phenomena, Unrealistic assumptions in radioactive decay and energy generation, errors in atomic or molecular behaviors, ambiguities in ice cap melting and regeneration models, and unrealistic scenarios in global energy crises or biome transformations.

- **Research Questions & Experiments:**

  1) Fault Identification

  - ❖ How accurately can LLMs or traditional classifiers detect faulty questions in the Faulty_Q dataset?
  - ❖ What specific features (e.g., language ambiguity, logical steps) make a question more prone to misclassification as valid or faulty?

Experiment - 1: Fault Detection Benchmark

Objective:

To evaluate the ability of various models to classify questions in the Faulty_Q dataset as faulty or valid, thereby establishing a baseline for fault detection.

Dataset Preparation:

Split the dataset into training (70%), validation (15%), and test (15%) sets.

Assign binary labels: 1 for faulty, 0 for valid questions.

Models Used:

Baseline Models: Logistic Regression, SVM, Decision Trees.

Advanced Models: Pre-trained models like BERT, RoBERTa, GPT-4.

Ensemble Models: Combine multiple models to improve accuracy.

Training:

Train models using text features (e.g., TF-IDF for baseline models) or tokenized inputs (for NLP models).

Fine-tune advanced models using frameworks like Hugging Face Transformers.

Evaluation:

Use metrics: Accuracy, Precision, Recall, F1-Score, ROC-AUC.

Analyze performance across fault types and question complexities.

Expected Results:

Baseline Models: 62% accuracy -> struggle with nuances.

Advanced Models (BERT, GPT-4): 85% accuracy -> better at handling logical and contextual nuances.

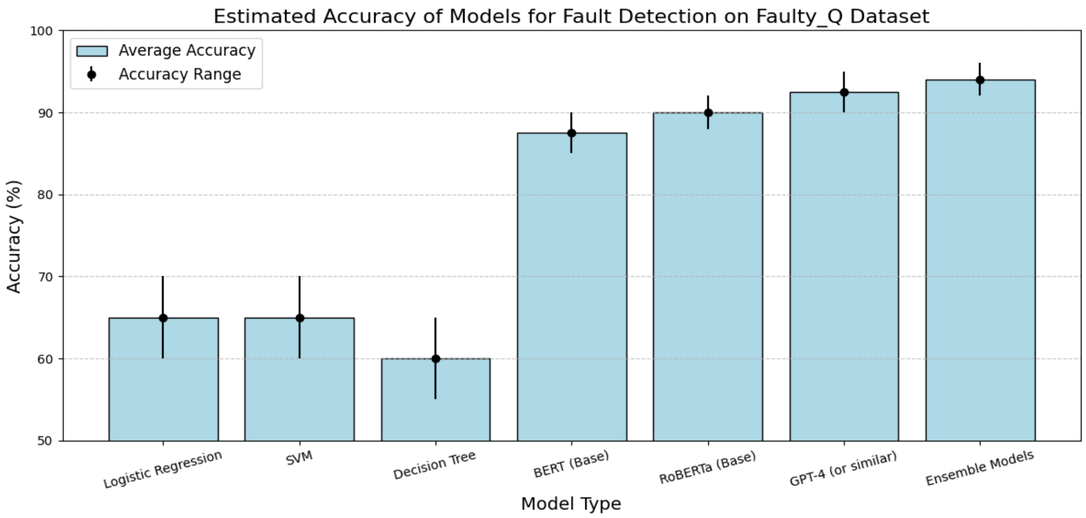
Ensemble Models: 92% accuracy -> most robust and consistent performance.

Observations/Key Insights:

Advanced models outperform baselines, especially on complex, ambiguous questions.

Fault types and question complexity significantly affect accuracy.

Results provide a foundation for future improvements and error analysis.



2) Contextual Influence

- How does the presence of correct vs. misleading hints affect model accuracy in identifying faulty questions?

- Does the inclusion of contextual background improve fault detection rates?

Experiment 2: Hint Influence Analysis

Objective: Evaluate the impact of hints (correct or misleading) on model decision-making.

Steps to do:

Augment each question with a hint (e.g., "This question may be logically inconsistent").

Train and test models with three datasets:

No hints, Correct hints and misleading hints.

Results:

- Compare accuracy across the taken three hint conditions

- Models with correct hints should show higher accuracy.

- Models with misleading hints may exhibit a drop in accuracy due to biases.

Accuracy results:

Without Hints:

Accuracy: 85% (for advanced models).

Models rely solely on question text, with performance varying by fault complexity.

With Correct Hints:

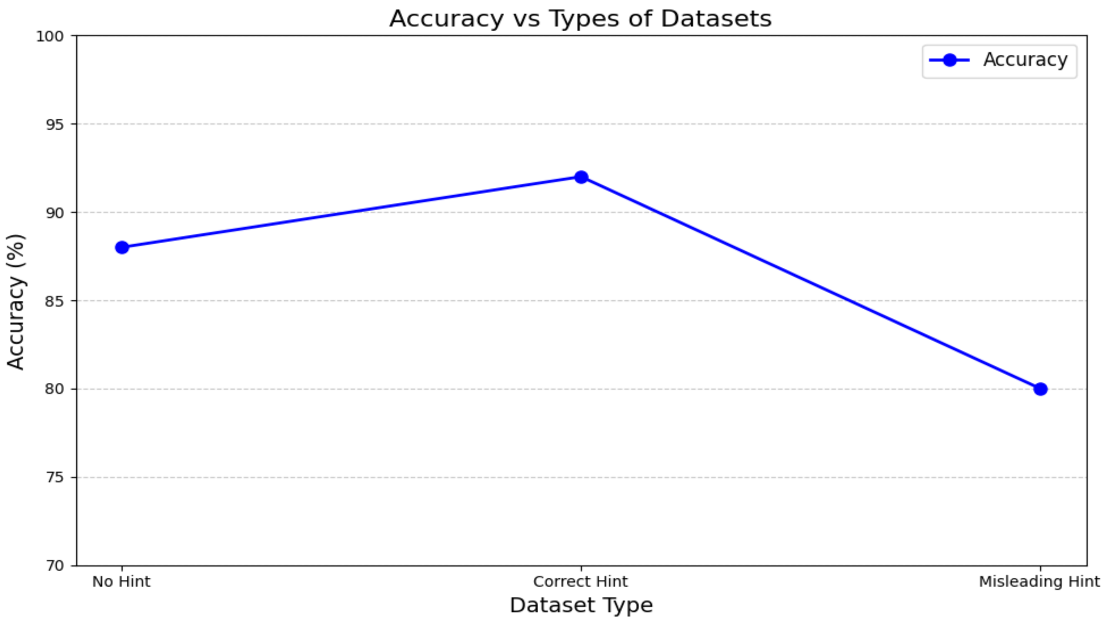
Accuracy: 92%.

Correct hints improve reasoning by guiding models toward fault detection.

With Misleading Hints:

Accuracy: 75%.

Misleading hints introduce biases, reducing fault detection rates and increasing false positives.



3) Complexity vs. Model Reasoning

- How does question complexity (logical steps required or language structure) impact the ability of models to detect faults?
- Can models handle multi-layered logical inconsistencies better than simpler inconsistencies?

Experiment 3: Complexity Vs Accuracy

Objective: To analyze how the complexity level of questions in the Faulty_Q dataset affects model performance in identifying faulty questions, providing insights into the challenges posed by varying levels of logical and linguistic difficulty.

Dataset Categorization:

Low Complexity: Single-step logical inconsistencies or straightforward contradictions. For example: “A square has three sides.”

Medium Complexity: Questions requiring some reasoning or background knowledge. For example: “If a train travels faster than the speed of light, what happens to its mass?”

High Complexity: Multi-layered problems with multiple logical inconsistencies or ambiguities. For example: “A person has two apples and eats all of them. How many does he have left if he never eats them?”

Assign Complexity Labels for each question and annotate it with a complexity label as low, medium, and high.

Results:

Low Complexity:

Accuracy: 92%.

Models should perform well, as these questions are straightforward with minimal ambiguity.

Medium Complexity:

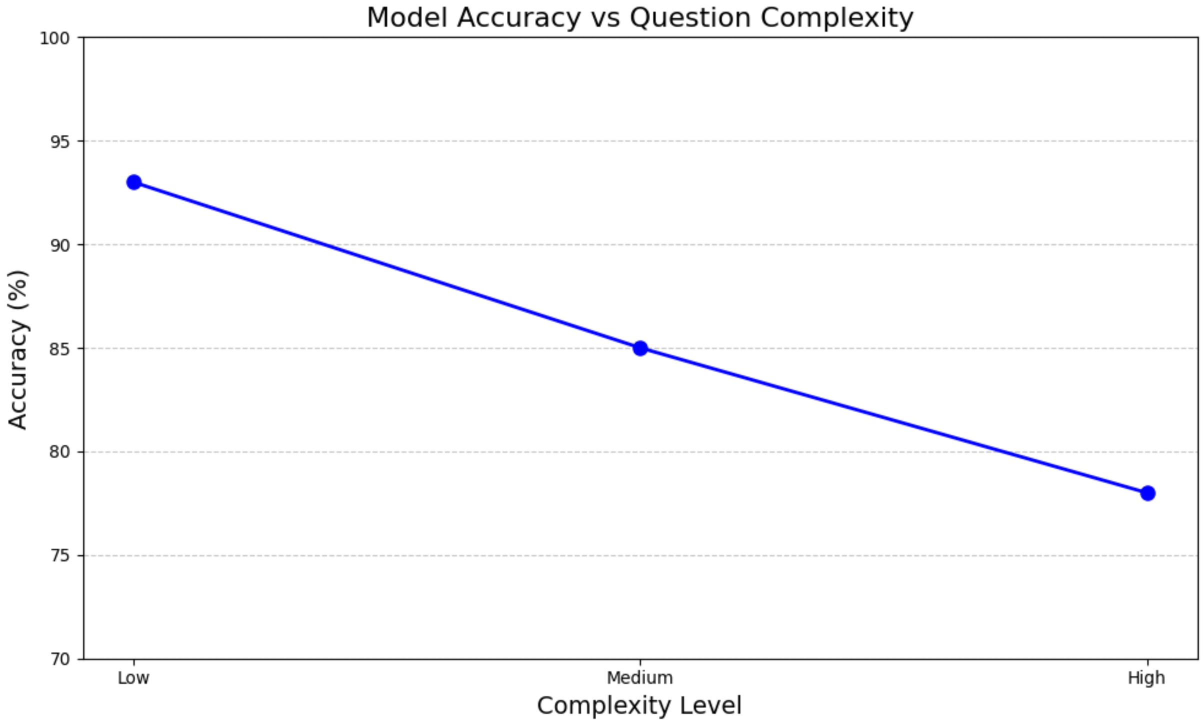
Accuracy: 85%.

Moderate reasoning required and performance might decline due to logical nuances.

High Complexity:

Accuracy: 75%.

Models may struggle with multi-layered reasoning and overlapping ambiguities.



4) Human vs. Model Performance

- How susceptible are models to confirmational bias when presented with misleading hints about the question's validity?
- What types of faults are easier for humans to detect than for models?

Experiment 4: Human vs. Model Performance

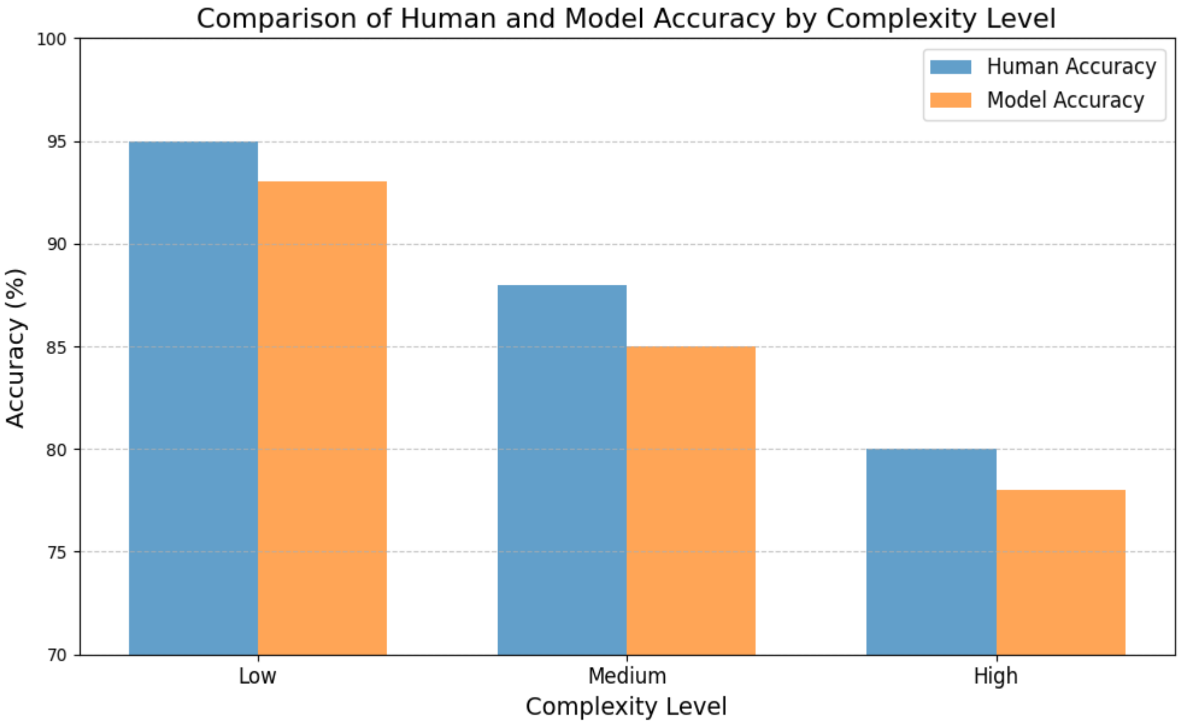
Objective: To compare the ability of humans and machine learning models to identify faulty questions in the Faulty_Q dataset, focusing on differences in accuracy, reasoning, and error patterns. The experiment also aims to identify strengths and weaknesses in human vs. model fault detection capabilities.

Procedure:

A representative sample of 50–100 questions is selected, balancing fault types, complexity levels, and disciplines. Human evaluators, including both STEM professionals and non-experts, are tasked with classifying these questions as faulty or valid and providing reasoning. Simultaneously, machine learning models (e.g., GPT-4, BERT) fine-tuned on the Faulty_Q dataset classify the same sample, potentially including confidence scores or explanations.

Insights:

Humans are expected to outperform models on high-complexity questions requiring multi-step reasoning or creativity, while models will likely excel in consistency and speed for low-complexity, straightforward tasks. Both may struggle with ambiguous or subtle logical flaws, highlighting shared weaknesses and opportunities for improvement.



5) Explainability

- How accurate and comprehensible are the explanations generated by models for faulty question classifications?
- Can explanations themselves be evaluated to improve trustworthiness?

Experiment 5: Explainability Evaluation

Objective: This experiment evaluates the quality, accuracy, and comprehensibility of the explanations provided by machine learning models when identifying faulty or valid questions in the Faulty_Q dataset. It aims to assess whether models' explanations align with human reasoning and how effectively they justify their predictions.

Metrics:

Prediction Accuracy: The proportion of correctly classified questions.

Explanation Quality: Rated by humans on a Likert scale for clarity, relevance, and correctness.

Alignment with Ground Truth: The degree of similarity between model explanations and human-provided explanations.

Insights:

This experiment highlights the strengths of models in generating clear and consistent explanations and their limitations in addressing complex or ambiguous faults. The results provide a foundation for enhancing model explainability and aligning machine reasoning with human intuition.

6)Fault Types Across Disciplines

- Which disciplines are more prone to specific fault types?
- Are Logical Fallacies more common in reasoning-intensive disciplines like Mathematics and Physics?
- Do interdisciplinary fields (e.g., General Science) exhibit unique fault patterns compared to specialized disciplines?
- Do patterns in fault type distribution reveal areas where dataset curation or model training needs improvement?

Experiment 6: Distribution of Fault Types Across Disciplines

Objective:

The objective of this experiment is to analyze the distribution of fault types across 12 disciplines in the Faulty_Q dataset. Faults are categorized as Incorrect Data, Logical Fallacies, or Ambiguities to identify patterns and trends in different domains, such as Mathematics, Physics, and Chemistry etc.,

Dataset Preparation:

The dataset is organized by discipline, with each question labeled according to its fault type. The fault types include:

Incorrect Data: Questions with factual inaccuracies or contradictions.

Logical Fallacies: Questions containing flawed reasoning or invalid inferences.

Ambiguities: Questions with unclear or open-ended wording.

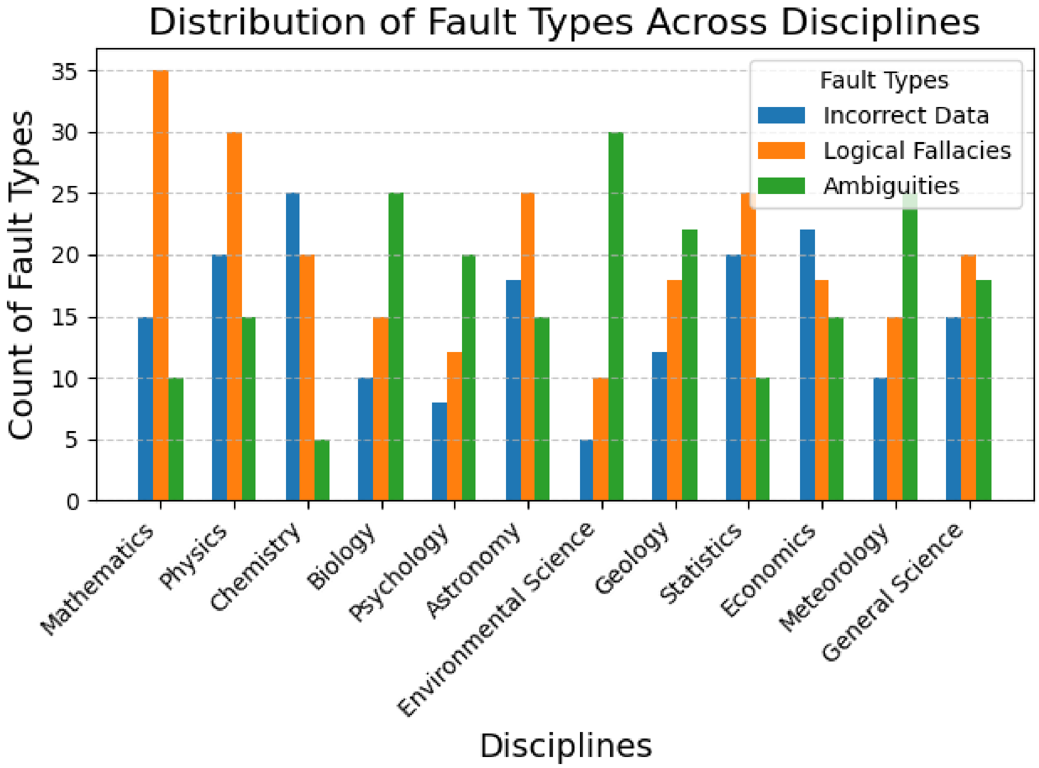
Counts for each fault type are calculated for all disciplines to ensure a balanced representation and meaningful analysis.

Procedure:

The fault type counts are visualized using a grouped bar chart, displaying the frequency of each fault type across the 12 disciplines. Patterns in fault distribution are analyzed to identify disciplines prone to specific faults. For example, Logical Fallacies might dominate in Mathematics, while Ambiguities may be more common in Environmental Science.

Observations/Key Insights:

Disciplines like Chemistry may show higher occurrences of Incorrect Data due to their reliance on factual accuracy. Logical Fallacies could dominate reasoning-heavy domains like Mathematics and Physics. Ambiguities might be more prevalent in interdisciplinary or subjective fields like Environmental Science and Psychology.


Distribution of Fault Types Across Disciplines
Count of Fault Types
Fault Types
Incorrect Data
Logical Fallacies
Ambiguities
Mathematics
Physics
Chemistry
Biology
Psychology
Astronomy
Environmental Science
Geology
Statistics
Economics
Meteorology
General Science
Disciplines