Question 1: How does the presence of numerical values in faulty questions influence LLMs' ability to detect logical flaws?

- Compare detection rates between qualitative vs quantitative faulty questions
- Analyze if LLMs are more likely to perform calculations without questioning validity when numbers are present
- Could reveal whether LLMs prioritize computation over logical validation

Experimental Setup and Methods: Numeric vs. Qualitative **Bayesian Reasoning Tests**:

- **Objective:** Explore whether presenting numeric information within a Bayesian reasoning framework elicits stronger premise evaluation. For instance, present probability-based questions that hinge on flawed base rates and see if LLMs differentiate between plausible and implausible numerical probabilities.
- **Potential Insight:** This could uncover whether certain problem structures (probabilistic rather than arithmetic) push LLMs from raw computation to the evaluation of underlying assumptions, thus bridging the gap between numeric engagement and logical scrutiny.

Methods

1. **Response Analysis Framework** I analyzed LLM responses across four key dimensions:

   a) **Probability Calculation Detection**

   - Indicators: 'p(', 'probability of', 'chance of', 'likelihood of', '%', 'percent'
   - Purpose: Identify when LLM attempts mathematical probability computations

   b) **Bayesian Concept Usage**

   - Indicators: 'prior', 'posterior', 'conditional probability', 'given that'
   - Purpose: Detect application of Bayesian reasoning principles

   c) **Constraint Checking**

   - Indicators: 'cannot exceed', 'impossible', 'must be between', 'valid range'
   - Purpose: Identify when LLM validates probability constraints

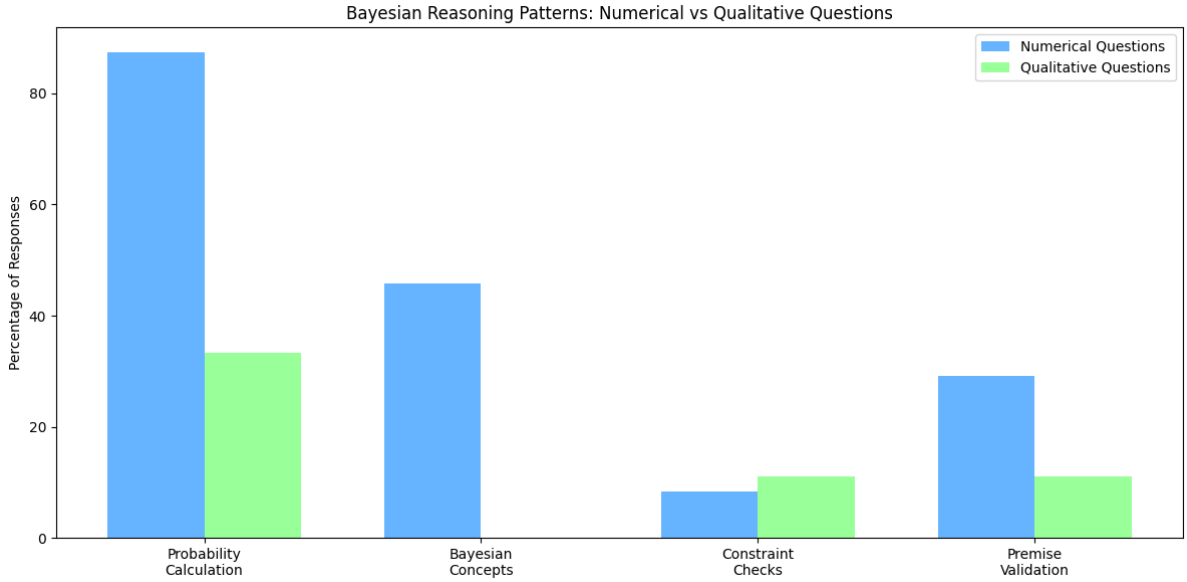
   d) **Premise Validation**

   - Indicators: 'assumption', 'premise', 'valid', 'invalid', 'contradiction'
   - Purpose: Detect critical evaluation of question premises

Result

```
=== Bayesian Reasoning Analysis ===

Total Questions: 166
Numerical Questions: 87 (52.4%)
```

```
Qualitative Questions: 79 (47.6%)

=== Response Patterns ===

Numerical Questions:
- Prob Calculation: 87.5%
- Bayes Concepts: 45.8%
- Constraint Checks: 8.3%
- Premise Validation: 29.2%

Qualitative Questions:
- Prob Calculation: 33.3%
- Bayes Concepts: 0.0%
- Constraint Checks: 11.1%
- Premise Validation: 11.1%
```



1. **Increased Propensity for Probability Calculations with Numbers:**
    - **Numerical Questions:** 87.5% of responses attempted probability calculations, reflecting the LLM's confidence in dealing with explicit numeric values.
    - **Qualitative Questions:** Only 33.3% attempted probability calculations, indicating that in the absence of numbers, the model is much less inclined to quantify uncertainty or engage in formal probabilistic reasoning.

    This disparity suggests that the LLM relies on numeric cues to trigger quantitative reasoning processes.

2. **Bayesian Concept Usage Tied to Numeric Cues:**
    - **Numerical Questions:** 45.8% of responses used Bayesian concepts such as prior and posterior probabilities.
    - **Qualitative Questions:** 0.0% usage of Bayesian concepts shows that when no numbers are provided, the model struggles or neglects to apply underlying Bayesian logic.

The presence of specific values appears to unlock the LLM's capacity (or inclination) to frame the problem in Bayesian terms, whereas qualitative frames fail to activate these more advanced reasoning tools.

3. **Constraint Checks Remain Low Regardless of Format:**
    - **Numerical Questions:** Constraint checks (e.g., verifying probabilities sum to 1 or identifying logical inconsistencies) appear in only 8.3% of responses.
    - **Qualitative Questions:** Constraint checks are also rare at 11.1%.

    Even when the LLM performs calculations, it seldom verifies whether results conform to probabilistic principles or logical constraints, underscoring a critical gap in model rigor.

4. **Premise Validation Slightly Better with Numbers, Still Weak Overall:**
    - **Numerical Questions:** Premise validation occurs at 29.2%, higher than in previous scenarios without numbers, suggesting that numerical detail nudges the model slightly toward scrutinizing assumptions.
    - **Qualitative Questions:** A mere 11.1% premise validation highlights that without numeric prompts, the model rarely questions the underlying scenario.

    Though numbers encourage some premise reflection, both figures indicate a persistent shortfall in the model's ability to challenge and validate foundational assumptions.

Conclusions:

- **Numeric Data as a Trigger for Bayesian Reasoning:** The presence of explicit numbers appears essential for the LLM to engage in probabilistic thinking and apply Bayesian frameworks.
- **Computation Over Validation:** While numbers promote computational effort and the invocation of Bayesian concepts, they do not substantially improve the model's ability to verify logical constraints or validate premises.
- **Qualitative Difficulty:** In the absence of numeric values, the LLM's probabilistic reasoning and validation efforts drop markedly, suggesting that it relies heavily on numeric cues to structure its reasoning rather than adapting flexible, abstract Bayesian reasoning in a qualitative context.

Potential Feature continue work: Dual-Mode Prompting Experiments (Calculation-First vs. Validation-First) → Compare LLM performance under two distinct prompting regimes: one that encourages immediate calculation and one that instructs the model to first validate premises before performing any numeric operations.

Question 2 Research Questions: Question vs. Response Length:

> Does the length of the flawed question prompt correlate with the length of the LLM's response, even though the LLM fails to detect the flaw?

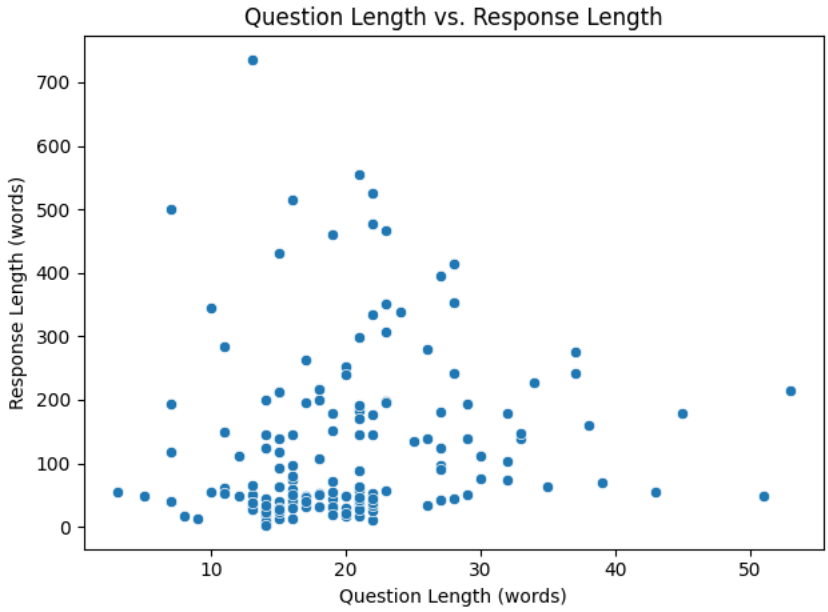
Experimental Approaches:

Length Correlation Analysis:

- **Method:**
  - Measure the word count of each flawed question.
  - Measure the word count of the LLM's response to that question.
  - Compute the Pearson correlation coefficient to determine if longer questions prompt longer responses.

Results:

```
Correlation between question length and response length:
0.12779640817687016 p-value: 0.10083171483676898
```



1. **Length Correlation Result:**
   - **Correlation between question length and response length:** 0.1278
   - **p-value:** 0.1008

   Interpretation:
   The correlation is small (0.128) and not statistically significant at a conventional alpha level ($p = 0.1008 > 0.05$). This suggests there is no strong linear relationship between how long the flawed question is and how long the LLM's response will be. In other words, making the question longer does not reliably lead to a longer response, at least not in a statistically robust manner.

Question 3 Domain-Specific Terminology Usage by Discipline**:**
How does domain-specificity impact LLMs' ability to detect faulty questions across different scientific disciplines?

- Compare detection rates across fields like physics, biology, chemistry, math
- Investigate whether LLMs show stronger performance in certain scientific domains
- Could reveal whether the models' training data has domain-specific biases

Experiment
Method:

- Define lists of domain-specific keywords for each discipline.
- Count how many such terms appear in each LLM response.
- Calculate the average frequency of these terms per discipline to assess how "domain-aware" or "domain-engaged" the LLM appears when confronted with flawed questions.

Results:

```
Average domain term usage by discipline:
     Discipline  domain_term_count
0  Astrophysics           0.250000
1           Bio           0.000000
2     Chemistry           0.214286
3          Math           0.071429
4       Physics           0.370370
5    Statistics           0.000000
```

  - Physics responses contain the highest average number of domain-specific terms (0.370370), suggesting the model uses more specialized or technical vocabulary when dealing with physics-related flawed questions.
  - Astrophysics (0.250000) and Chemistry (0.214286) also show relatively higher engagement with domain-specific terms.
  - Biology and Statistics show no usage of the predefined domain keywords (0.000000), which could mean the LLM responded in more general terms or that the chosen keywords do not match how the LLM discusses these fields.

Key findings:

- The LLM's response length is not closely tied to the question length in these flawed scenarios, indicating that complexity or verbosity of the question does not necessarily prompt lengthier responses.
- The presence of domain-specific keywords in responses varies significantly by discipline, with Physics and Astrophysics showing higher engagement. This could imply that the LLM tries to leverage more specialized terminology in certain fields, even when failing to detect the underlying flaw, while remaining relatively generic in fields like Biology and Statistics.

Question 4 Can we cluster LLM responses to flawed questions to uncover underlying response patterns, and do these patterns correlate with certain disciplines or latent topics?

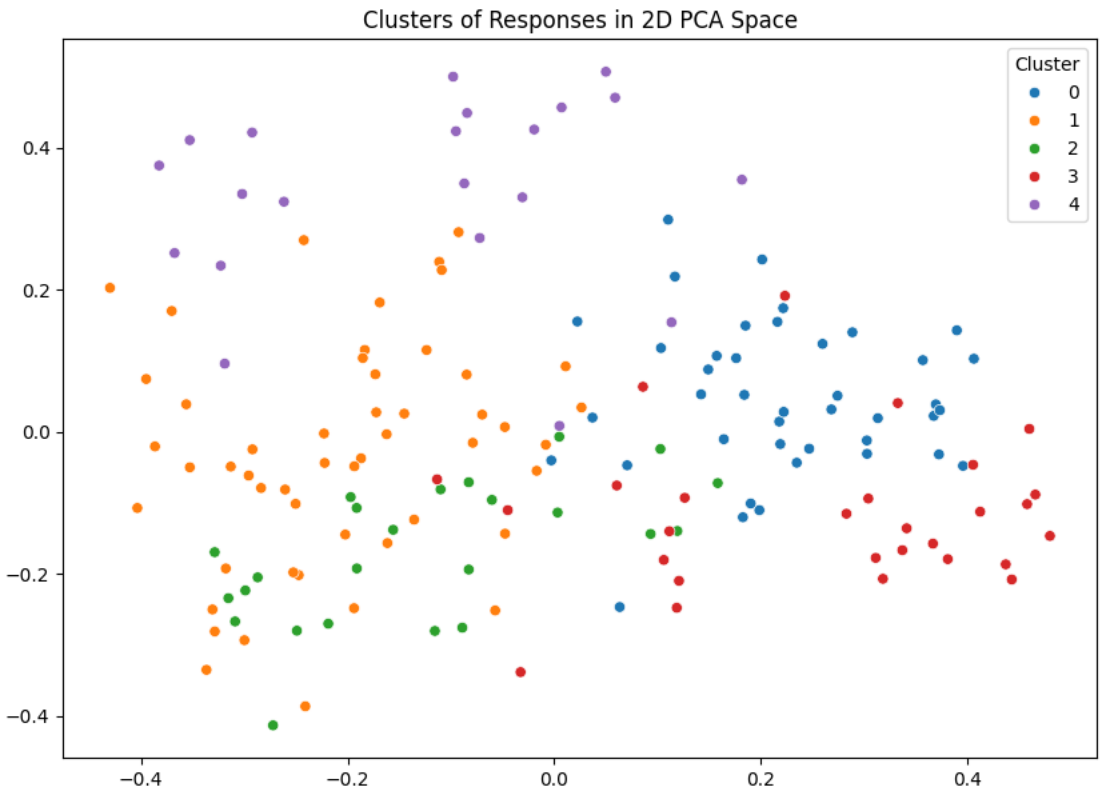
Experiment

- **Embedding and Clustering:**
  - Used a sentence transformer model to embed LLM responses into numeric vectors.
  - Applied K-Means clustering on the embeddings to group similar responses.
  - Examined the distribution of responses across clusters and how those clusters relate to the discipline of each question.

- **Topic Modeling (LDA):**
  - Converted responses into TF-IDF representations.
  - Applied Latent Dirichlet Allocation (LDA) to identify latent topics in the responses.
  - Investigated which topics appear most frequently and how their distribution relates to the discipline of each question.

Result

```
Cluster Distribution:
cluster
1    51
0    41
3    29
2    24
4    21
Name: count, dtype: int64
```




```
Discipline vs. Cluster:
cluster        0   1   2   3   4
Discipline
Astrophysics   2   9   0   0  17
Bio            3   8  12   4   0
Chemistry      3  16   9   0   0
Math          25   2   0   1   0
Physics        7  11   3   2   4
Statistics     1   5   0  22   0
Topic 0: pieces, mode, orbital, 16, number, work, apples, right, com,
different
Topic 1: cm, ph, 10, probability, 14, number, 100, dark, negative, years
Topic 2: cell, 12, probability, water, candidate, relationship, linear,
333, female, methane
Topic 3: 10, force, temperature, probability, metaphase, days, number,
light, position, water
Topic 4: miles, train, distance, time, distribution, square, minutes,
number, hazard, half

Topic Distribution:
topic
3    38
0    36
1    32
2    31
4    29
Name: count, dtype: int64

Discipline vs. Topic:
topic         0  1  2  3  4
Discipline
Astrophysics  4  6  6  9  3
Bio           7  5  4  8  3
```

Chemistry	7	7	7	4	3
Math	7	6	3	4	8
Physics	6	4	4	7	6
Statistics	5	4	7	6	6

Key Findings:

- **Clustering:** The formation of distinct clusters associated with certain disciplines suggests that the LLM's style or approach to answering flawed questions may vary by domain.
- **Topics:** While topics are more evenly spread across disciplines, they highlight recurrent themes (e.g., numerical reasoning, probabilities, measurements) the LLM uses in responses, even when failing to detect flawed premises.

Potential deeper research followed by question 4

- **Idea:** Move beyond broad flaw categories (e.g., "impossible scenario") and develop a more nuanced taxonomy of errors (e.g., "violates fundamental constant," "contradicts established equilibrium condition," "assumes nonexistent particle").
- **Goal:** Determine if certain clusters or topics align strongly with highly specific **flaw types**, revealing subtler reasoning failures.