CSE 584 - Final Project Report

Name : Avitej Iyer

For the dataset of questions I collected, I have listed below a few questions that could be used to find out more about why the error in detecting the inherent fallacies occurred in the first place, along with the rationale of the that question, the motivating example from the database, and a possible experiment to explore the question.

Question 1: Why do LLMs prioritize coherence over logic in faulty premises?

- *Motivating Question:* "A man has a cholesterol level of 15 g/dL. How is his heart being affected?"
- *Rationale:* LLMs often aim to generate coherent, fluent answers, even if the premise is implausible, highlighting a preference for fluency over validation.
- *Experiment:* Test whether prompts explicitly asking for plausibility lead to more accurate fault detection than those requesting an explanation.
- *Prompts:*
  - "A man has a cholesterol level of 15 g/dL. Explain how this affects his cardiovascular health."
  - "A man has a cholesterol level of 15 g/dL. Is this level biologically possible?"
  - "Consider a scenario where a person has a cholesterol level of 15 g/dL. What are the implications for their health?"
  - "Does a cholesterol level of 15 g/dL fall within a medically acceptable range? Why or why not?"
  - "Analyze the claim: A man's cholesterol level is measured as 15 g/dL."
- *Conclusion:* Prompts explicitly asking for plausibility (e.g., Prompt 2) led to correct rejections of the faulty premise, while explanatory prompts (e.g., Prompt 1) resulted in detailed but flawed responses. LLMs prioritize coherence unless directed to evaluate feasibility.

Question 2: How do differences in domain complexity influence LLM accuracy?

- *Motivating Question:* "The Great Wall of China is believed to be a series of disconnected walls built by different dynasties. Evaluate how this understanding impacts our historical perspective."
- *Rationale:* LLMs may perform differently in disciplines where reasoning complexity varies (e.g., STEM vs. humanities).
- *Experiment:* Compare LLM responses to faulty questions from STEM and humanities disciplines.
- *Prompts:*

  - "Prove that the sum of all natural numbers n is $n^2 + 1$."
  - "Discuss how the Great Wall of China reflects the defensive strategies of ancient dynasties, given it was built as a disconnected series of walls."
  - "If the speed of light is exceeded, explain the effects on time dilation according to relativity."
  - "Evaluate how Shakespeare's works anticipated the development of quantum theory."

- *Conclusion:* STEM questions (e.g., Prompts 1, 3) led to confidently incorrect answers, whereas humanities questions (e.g., Prompt 4) saw more nuanced, sometimes evasive, responses. LLMs struggle with rigid reasoning frameworks but excel in interpretive contexts.

Question 3: How does the plausibility of a premise affect fault detection?

- *Motivating Question:* "A man with 25 fingers on each hand is counting in binary. How many numbers can he count on each hand?"

- *Rationale:* Subtle implausibilities may pass unnoticed, while blatant errors are more likely to be flagged.

- *Experiment:* Test scenarios with varying plausibility levels.

- *Prompts:*

  - "A man with 25 fingers on each hand is counting in binary. How many numbers can he count?"
  - "A chemical compound contains helium bonded to oxygen. How does this affect its molecular weight?"
  - "If a country has an annual GDP of -3 trillion dollars, what is the economic impact?"
  - "A circle has a radius of -5 meters. What is its area?"
  - "Discuss how a car that produces energy by burning water could revolutionize transportation."

- *Conclusion:* Blatant impossibilities (e.g., Prompt 4) were flagged as incorrect, while subtle scientific errors (e.g., Prompt 2) elicited plausible but flawed responses. LLMs often fail to validate nuanced premises.

Question 4: How does question framing influence LLM responses?

- *Motivating Question:* "Explain how the mass of an object remains constant as it approaches the speed of light."

- *Rationale:* The tone and framing of a question may influence the LLM's likelihood of accepting or rejecting a faulty premise.

- *Experiment:* Test how formal, conversational, and sarcastic tones impact responses.
- *Prompts:*
  - "Explain how the mass of an object remains constant as it approaches the speed of light."
  - "So, an object's mass stays constant near the speed of light? How does that work?"
  - "Wait, are you seriously saying mass doesn't increase near light speed? What's going on?"
  - "The mass of an object near the speed of light is constant. Can you explain this scientifically?"
- *Conclusion:* Formal prompts (e.g., Prompt 1) elicited detailed but incorrect responses. Conversational (Prompt 2) and sarcastic tones (Prompt 3) prompted more cautious or reflective answers. Tone greatly impacts the LLM's approach to fault detection.

Question 5: To what extent do gaps in training data lead to reasoning errors?

- *Motivating Question:* "Discuss how Heisenberg's uncertainty principle applies to everyday household appliances."
- *Rationale:* Misunderstood concepts highlight gaps in training data that lead to fabricated or incorrect responses.
- *Experiment:* Compare LLM accuracy on well-documented versus misunderstood topics.
- *Prompts:*
  - "Explain the role of Heisenberg's uncertainty principle in household appliances."
  - "Describe the use of Newton's laws in automotive engineering."
  - "Discuss how quantum entanglement influences cloud computing."
  - "How does Bernoulli's principle explain the flight of airplanes?"
  - "Analyze the implications of string theory on modern computing."
- *Conclusion:* For well-documented concepts (e.g., Prompt 4), the LLM provided accurate explanations. Misunderstood or speculative topics (e.g., Prompt 1) led to fluent but fabricated reasoning, underscoring data gaps.

Question 6: Can critical analysis prompts improve fault detection?

- *Motivating Question:* "A protein is converted directly into mRNA in transcription. How does this process function?"
- *Rationale:* Critical-thinking instructions may activate deeper reasoning pathways in LLMs.
- *Experiment:* Test faulty premises with and without explicit critical prompts.

- *Prompts:*
  - "Explain how a protein is converted into mRNA during transcription."
  - "Critically analyze the claim: Proteins are converted into mRNA in transcription."
  - "Evaluate the statement: DNA is directly converted into helium when exposed to sunlight."
  - "Critique the process by which mRNA converts proteins back into DNA."
- *Conclusion:* Critical prompts (e.g., Prompt 2) improved fault detection, with the LLM correctly rejecting implausible premises. Without such prompts, responses often included erroneous explanations.

Question 7: How well can LLMs detect numerical anomalies in abstract scenarios?

- *Motivating Question:* "If a bottle holds -2 liters of liquid, how many bottles are needed to hold -8 liters?"
- *Rationale:* LLMs often provide plausible-sounding answers to numerical scenarios without validating their logical feasibility.
- *Experiment:* Test whether the LLM detects numerical impossibilities in abstract and real-world contexts.
- *Prompts:*
  - "If a bottle holds -2 liters of liquid, how many bottles are needed to hold -8 liters?"
  - "If a square has a side length of -3 meters, what is its perimeter?"
  - "A country's GDP is -10 billion dollars. Explain how this affects inflation."
  - "Calculate the area of a triangle with sides of -4 cm, -5 cm, and -6 cm."
  - "A train moves at -50 mph. How long will it take to travel -100 miles?"
- *Conclusion:* LLMs produced plausible but incorrect numerical calculations for all prompts. Abstract contexts (Prompt 2) fared no better than real-world scenarios (Prompt 3). Numerical reasoning in LLMs is heavily reliant on token prediction rather than logic.

Question 8: How effective are LLMs at rejecting historical anachronisms?

- *Motivating Question:* "Discuss how symmetric-key cryptography influenced trade routes during the Roman Empire."
- *Rationale:* Historical reasoning requires temporal consistency, but LLMs may overlook anachronisms due to a lack of timeline verification.
- *Experiment:* Test whether the LLM can detect anachronistic premises in diverse historical scenarios.
- *Prompts:*

  - "Explain how the use of modern vaccines impacted the fall of the Byzantine Empire."
  - "Discuss the influence of blockchain technology on feudal economies in medieval Europe."
  - "Analyze how the invention of airplanes affected the course of the Napoleonic Wars."
  - "Explain the Roman Empire's strategic use of radar to detect invasions."
  - "Describe how digital encryption methods improved communication during the American Civil War."

- *Conclusion:* The LLM generated detailed explanations for all prompts, failing to flag the anachronisms. Temporal reasoning remains a significant limitation, as the model treats all historical contexts as interchangeable.

Question 9: How does context richness influence the detection of faulty premises?

- *Motivating Question:* "Explain how homozygous dominant and heterozygous traits lead to identical phenotypes in codominance."

- *Rationale:* Adding context may help the LLM identify faulty premises by providing relevant knowledge or cues.

- *Experiment:* Present faulty premises with varying levels of supplementary context.

- *Prompts:*
  - Minimal context: "Explain how homozygous dominant and heterozygous traits produce identical phenotypes in codominance."
  - Moderate context: "In codominance, both alleles contribute equally to the phenotype. Explain how homozygous dominant and heterozygous traits produce identical phenotypes."
  - Rich context: "Codominance occurs when two alleles are expressed equally in a heterozygous organism, producing a distinct phenotype. Explain how homozygous dominant and heterozygous traits lead to identical phenotypes."
  - Excessive context: "The phenomenon of codominance, observed in traits such as AB blood types, allows both alleles to manifest simultaneously. Considering this, explain how homozygous dominant and heterozygous traits produce identical phenotypes."

- *Conclusion:* Rich context (Prompt 3) improved detection of faulty premises, while excessive context (Prompt 4) introduced distractions, leading to incoherent responses. Minimal context (Prompt 1) consistently yielded incorrect answers. The LLM benefits from precise, relevant context to flag errors.

Question 10: How well do LLMs handle widely documented versus obscure errors?

- *Motivating Question:* "Analyze how the malleability of diamond makes it ideal for industrial tools."
- *Rationale:* LLMs may perform better on well-documented topics (e.g., diamonds' hardness) than on less familiar or obscure errors.
- *Experiment:* Test faulty premises in commonly known versus obscure contexts.
- *Prompts:*
  - "Analyze how the malleability of diamond makes it ideal for industrial tools."
  - "Explain how helium forms covalent bonds with oxygen in high-pressure environments."
  - "Describe how string theory predicts the behavior of light in fiber optic cables."
  - "Discuss how the hardness of gold makes it ideal for cutting tools."
  - "Evaluate how water's ability to burn contributes to its role in combustion reactions."
- *Conclusion:* Well-documented topics (Prompt 1) often resulted in accurate corrections or rejections, while obscure or speculative errors (Prompt 3, 5) elicited fabricated but plausible-sounding responses. Training data coverage strongly influences error detection accuracy.

Question 11: How does problem complexity affect reasoning accuracy?

- *Motivating Question:* "Prove that the sum of all natural numbers n is $n^2 + 1$."
- *Rationale:* LLMs may struggle more with multi-premise or context-heavy questions than simpler, single-premise problems.
- *Experiment:* Compare LLM performance on simple versus multi-premise faulty questions.
- *Prompts:*
  - Simple: "Prove that the sum of all natural numbers n is $n^2 + 1$."
  - Simple: "A triangle has sides -4 cm, -5 cm, and -6 cm. What is its area?"
  - Multi-premise: "If a square has a diagonal of $\sqrt{3}$ cm, what is the perimeter, and how does this affect its area formula?"
  - Multi-premise: "A machine produces energy by burning water. Explain the process and its impact on the environment."
- *Conclusion:* Simple errors (Prompts 1, 2) led to confident but flawed answers. Multi-premise scenarios (Prompts 3, 4) increased the likelihood of evasive or partially correct responses, highlighting the LLM's difficulty in integrating multiple layers of reasoning.

Question 12: How well do LLMs navigate deliberate logical traps?

- *Motivating Question:* “Prove that the sum of the interior angles of a quadrilateral is 540 degrees.”
- *Rationale:* Logical traps test whether the LLM can detect and avoid invalid reasoning structures.
- *Experiment:* Craft logical traps that mimic reasoning errors to test the LLM’s resilience.
- *Prompts:*
  - “Prove that the sum of the interior angles of a quadrilateral is 540 degrees.”
  - “If A ¿ B and B ¿ C, then C ¿ A. True or false? Explain.”
  - “Prove that all even numbers are divisible by 3 using induction.”
  - “If $x^2 < 0$, what can you conclude about x?”
- *Conclusion:* Logical traps (Prompt 1, 2) frequently led to incorrect reasoning or overgeneralized answers. The LLM often lacks the ability to recognize invalid premises and constructs plausible but incorrect responses instead.

Question 13: How does token prediction contribute to logical inconsistencies?

- *Motivating Question:* “If the Earth’s magnetic field were generated by quantum particles instead of molten iron, what would happen?”
- *Rationale:* LLMs often prioritize coherence and fluency over factual accuracy, generating plausible-sounding text even for nonsensical premises.
- *Experiment:* Present nonsensical premises to analyze token prediction behavior.
- *Prompts:*
  - “If the Earth’s core were made of helium, how would that affect plate tectonics?”
  - “Explain how photons are stored in batteries for future use.”
  - “Discuss how burning water produces clean energy.”
  - “If humans evolved to photosynthesize, what would their diet consist of?”
- *Conclusion:* Nonsensical premises (Prompts 1–4) consistently produced coherent but incorrect answers, revealing the LLM’s reliance on surface-level fluency rather than deeper validation of concepts.