ML CSE584 Final Project

Balaji Udayagiri gbu5048

December 7, 2024

1 Introduction

The advent of advanced language models has ushered in a new era of artificial intelligence (AI), enabling machines to perform complex reasoning and problem-solving tasks across diverse domains. From scientific inquiry to creative exploration, these models have demonstrated remarkable potential in interpreting and generating human-like responses. However, their limitations in contextual understanding, domain-specific reasoning, and logical consistency necessitate a comprehensive evaluation to better understand their capabilities and areas for improvement.

This report investigates the performance of a state-of-the-art language model across five distinct domains: chess analysis, biology, physics, mathematics, and chemistry. These domains were chosen for their diversity, requiring a combination of logical reasoning, domain-specific expertise, and contextual awareness. By analyzing the model's responses to carefully designed prompts in each domain, this report aims to provide a holistic understanding of its strengths, weaknesses, and potential applications.

Objectives of the Study

The primary objectives of this report are:

- To assess the model's ability to perform structured and unstructured reasoning across diverse disciplines.
- To identify common patterns of success and failure in its responses, highlighting both strengths and limitations.
- To propose targeted improvements for enhancing the model's contextual understanding and reasoning capabilities.

2 Chess Analysis

2.1 Research Question 1: Validity of Chess Positions

Question: Can the model detect invalid chess positions, such as those where pieces are placed on the same square?

Experiment:

- Asked if two pieces were placed on the same square.

Outcome:

- Initially, the model failed to recognize when two pieces occupied the same square in some cases.
- Upon follow-up prompts, the model acknowledged the error and provided accurate validation.

—

2.2 Research Question 2: Detection of Pawn Promotion

Question: Can the model detect when a pawn reaches the last rank and should be promoted?

Experiment:

- Presented a position with a pawn on the last rank.

Outcome:

- Initially, the model did not identify the need for pawn promotion.
- Upon follow-up prompts, it recognized the issue and suggested promotion.

—

2.3 Research Question 3: Check Detection

Question: Can the model detect when a player's king is in check?

Experiment:

- Asked whether the white king was in check in a specific position.

Outcome:

- The model identified check in some cases.
- However, it failed to evaluate check for both kings in the same position without explicit prompting.

—

2.4 Research Question 4: Validity of Chess Moves and Best Move Suggestion

Question: Can the model suggest the best move for a given side and validate its legality?

Experiment:

- Presented a position and asked for the best move for black.

Outcome:

- The model successfully suggested moves in simple positions.
- For complex positions, it sometimes failed to identify the best move involving both tactical and strategic considerations.

—

2.5 Observations and Analysis

- The model's ability to validate the legality of a chess position depends heavily on the context and the specific position presented.
- Follow-up prompts often lead to more accurate analysis.
- Critical errors, such as failing to detect check for both kings, suggest areas for improvement.

—

3 Biology Analysis

3.1 Does Question Length Matter?

Question: Does the length of a question impact the model's ability to detect biologically unrealistic magnesium concentrations?

Experiment:

- A short question: *"A certain blood sample has magnesium of 8 mg/mL. Is this realistic?"*
- A long, detailed question: *"Using a spectrophotometric method, calculate the magnesium concentration in blood based on given absorbance and other parameters."*

Outcome:

- For the short question, the model correctly identified the magnesium level as unrealistic.
- For the long question, the model performed the calculation but failed to flag the result as biologically implausible.

—

3.2 Does Presenting Concentrations Explicitly vs. Asking for Calculations Matter?

Question: Does explicitly stating the concentration versus embedding it in a calculation affect the model's ability to detect unrealistic values?

Experiment:

- Explicit concentration: *"The blood sample contains 8 mg/mL magnesium. Is this realistic?"*
- Calculation-based concentration: *"Calculate the magnesium concentration based on given absorbance and ratios."*

Outcome:

- When the concentration was explicitly stated, the model correctly identified the level as unrealistic.
- When asked to calculate the concentration, the model performed the calculation but did not evaluate its plausibility.

—

3.3 Does the Compound in Question Matter?

Question: Does the type of analyte (e.g., magnesium, glucose, arsenic) affect the model's ability to identify biologically plausible levels?

Experiment:

- Magnesium-related question: *"The blood sample contains 8 mg/mL magnesium. Is this realistic?"*
- Glucose-related question: *"The blood sample contains 8 mg/mL glucose. Is this realistic?"*
- Arsenic-related question: *"A blood sample contains 10 mg/mL arsenic. Is this realistic?"*

Outcome:

- For magnesium, the model identified unrealistic concentrations when explicitly prompted.
- For glucose, the model failed to flag biologically unrealistic levels.
- For arsenic, the model did not recognize the toxicity of high levels in blood.

—

3.4 Observations and Analysis

- Short, explicit questions lead to better anomaly detection than long, detailed ones.
- Explicitly stated concentrations improve anomaly detection.
- The compound type significantly affects responses. The model performs better with magnesium but struggles with glucose and arsenic.

4 Physics Analysis

This section explores the experiments conducted to evaluate the model's ability to analyze physical scenarios involving electric fields and nuclear fission. It investigates how variations in the question format, parameters, and explicitness of details influence the model's responses.

4.1 Does the Model Recognize Non-Conservative Fields?

Question: Can the model identify that the given electric field is non-conservative and invalidate the work-done calculation?

Experiment:

- Presented a non-conservative electric field: *"The Electric Field is given by:* $E_x = x_{coeff} perm[0]^{x_{power}} i + E_y = y_{coeff} perm[1]^{y_{power}} j + E_z = z_{coeff} perm[2]^{z_{power}} k$*. A charge is moved from (0,* y_{coord}*, 0) to (0, 0,* z_{coord}*). What is the work done?"*

- Asked the model to calculate the work done for multiple variations of non-conservative fields.

Outcome:

- The model assumed the field was conservative and performed the work-done calculation without checking the validity of the field.

- When explicitly prompted with a follow-up question about validity, the model calculated the curl of the field, identified it as non-zero, and correctly flagged the question as invalid.

—

4.2 Does Changing the Field Parameters Affect Responses?

Question: Does the model's ability to detect non-conservative fields improve with variations in the electric field's parameters?

Experiment:

- Changed the coefficients and powers in the electric field components for different test cases.

- Examples of varied fields: $E = x_{coeff} perm[0]^{x_{power}} i + y_{coeff} perm[1]^{y_{power}} j + z_{coeff} perm[2]^{z_{power}} k$ $E = x_{coeff} . perm[0]^{x_{p} ower} i + y_{coeff} . perm[1]^{y_{power}} j + z_{coeff} . perm[2]^{z_{p} ower} k$

Outcome:

- Changing the field parameters did not influence the model's initial response. It consistently assumed the field was conservative.

- Follow-up questions about field validity led to the correct realization that the curl was non-zero and the field was non-conservative.

—

4.3 Does the Model Identify Physically Impossible Speeds?

Question: Can the model recognize when the calculated speed of a particle exceeds the speed of light during nuclear fission scenarios?

Experiment:

- Presented a question where a large body becomes stationary after transferring momentum to a small particle undergoing fission: *"The momentum of a body weighing large_mass kg is large_p kgs^{-1}. It strikes a little item having a mass of little_mass milligrams, which undergoes nuclear fission. The particle divides into two parts in the mass ratio of mass_ratio, and the released energy causes the larger body to become stationary. How much energy is released during nuclear fission?"*

- Experimented with variations:

  - Explicitly mentioned that there is no loss of energy during the process.
  - Omitted details about the released energy causing the larger body to become stationary.
  - Added a sub-question asking to calculate the speed of the smaller particle.

Outcome:

- When the question explicitly mentioned that there was no energy loss and the large body became stationary, the model failed to recognize that the calculated speeds exceeded the speed of light.

- When these details were omitted, the model flagged the question as invalid, recognizing the physical implausibility of the scenario.

- When asked to calculate the speed of the smaller particle, the model correctly identified that the speed exceeded the speed of light and flagged the scenario as invalid.

4.4 Observations and Analysis

- The model assumes fields are conservative by default, leading to incorrect work-done calculations for non-conservative fields.

- Variations in electric field parameters or question framing do not change the model's initial assumption of conservativeness.

- In nuclear fission scenarios, the model fails to detect physically impossible speeds unless explicitly prompted to calculate them.
- Adding explicit checks (e.g., calculating speeds) or omitting critical assumptions prompts the model to recognize invalid scenarios.

5 Mathematics Analysis

This section explores the model's capabilities and limitations in solving mathematical problems, particularly in complex areas such as Fourier transforms, analytic continuation, and proving convergence or divergence. We also test the model's ability to identify incorrect premises and provide accurate reasoning.

5.1 Competence in Advanced Mathematics

The model demonstrates strong competence in complex mathematical domains, including Fourier analysis, Lebesgue integration, and analytic continuation. For example:

Experiment:

- Question: "I'm having trouble comprehending Plancherel's theorem while I study the Fourier transform. I am aware that Lebesgue integrals are used in the Fourier transform. Using Lebesgue integrals, I want to show Plancherel's theorem for the function $f(x) = \sqrt{1+x^2}$ after applying Parseval's finding."
- Outcome: The model correctly identified that the given function $f(x)$ is not suitable for demonstrating Plancherel's theorem using Parseval's identity. It provided an accurate explanation of why the function did not meet the required conditions.

—

5.2 Experiment: Can GPT Be Fooled with Incorrect Premises?

The model occasionally falters when asked to prove or disprove incorrect premises, particularly in subtle cases involving complex functions.

Example: Questions Involving Convergence and Integration

1. **Question:** Consider the function $f(x)$ defined as:

$$f(x) = \begin{cases} e^x \sin(1/x) & \text{if } x \neq 0, \\ 0 & \text{if } x = 0. \end{cases}$$

   Prove that the series $\sum f(1/n)$ converges. **Expected Answer:** It cannot be proved as the series diverges. **Outcome:** The model incorrectly provided a proof of convergence, failing to recognize that the series diverges due to the oscillatory behavior of $\sin(1/x)$.

2. **Question:** Prove that the integral of $|f(x)|$ from -1 to 1 is finite. **Expected Answer:** It cannot be proved, as the integral does not converge due to the oscillations of $\sin(1/x)$. **Outcome:** The model attempted to provide a solution but did not identify the integral's divergence.

3. **Question:** Consider the function $f(x) = \frac{x \sin(1/x)}{\sin x}$ (for $x \neq 0$) and $f(0) = 0$. Prove that $\lim_{x \to 0} f(x) = 0$. **Expected Answer:** It cannot be proved because $\sin(1/x)$ does not have a limit as $x \to 0$. **Outcome:** The model failed to recognize the invalidity of the limit proof and provided a faulty response.

—

Correct Behavior for Valid Problems: When the exponential e^x in the above examples is replaced by x, the questions become valid, and the model provides flawless proofs. This indicates that the model performs well with correctly framed problems but struggles with subtle divergence issues in oscillatory cases.

—

5.3 Response to Obvious Errors

The model generally performs well in identifying obvious mathematical errors. For example:

- When asked to prove that a convergent integral does not exist, the model correctly recognized that the integral existed and provided a valid counterargument.

—

5.4 Performance in Complex Scenarios

While the model excels in many advanced mathematical tasks, it occasionally fails in more nuanced or ambiguous scenarios.

Experiment:

- **Question:** "While trying to understand the similarity between Plancherel's and Parseval's theorems, show that I can't apply them on $\frac{1}{\sqrt{|x|\sqrt{x-1}}}$."

  Expected Answer: It is possible to demonstrate the applicability of these theorems to this function. **Outcome:** The model incorrectly stated that it could not demonstrate this, despite the question being valid.

—

5.5 Observations and Analysis

- The model demonstrates strong capabilities in advanced mathematical reasoning, particularly in Fourier analysis and integrals involving Lebesgue theory.
- It struggles with subtle issues in oscillatory functions or conditions where divergence is less apparent, leading to occasional faulty proofs.
- For clearly invalid or overly ambiguous premises, the model generally identifies the errors.
- Nuanced cases requiring careful analysis, such as specific convergence proofs, can occasionally lead to incorrect conclusions.

6 Chemistry Analysis

This section evaluates the model's reasoning and performance in solving calorimetry problems. The experiments focus on the model's assumptions, its ability to account for varying conditions, and its responsiveness to explicit prompts about missing information.

6.1 Assumptions in Calorimetry Problems

Question: Does the model consider all relevant physical conditions, such as pressure and phase transition points, when solving calorimetry problems?

Experiment:

- Presented the model with a calorimetry problem where key conditions, such as pressure, were not explicitly stated.
- Asked the model to calculate energy changes assuming phase transitions (e.g., melting and boiling).

Outcome:

- The model implicitly assumed standard conditions (1 atm pressure) and used typical values for melting and boiling points (e.g., 0°C and 100°C for water).
- It did not check for pressure variations or other conditions that could affect phase transition temperatures.
- When explicitly asked about the missing conditions, the model acknowledged that they were not provided, recognizing the need for additional data to ensure accurate calculations.

—

6.2 Explicit Prompts on Missing Conditions

Question: How does the model respond to explicit prompts about missing physical conditions in calorimetry problems?

Experiment:

- Presented the model with a calorimetry problem and then explicitly asked, "What assumptions have you made about pressure and phase transition points?"

Outcome:

- When prompted explicitly, the model recognized its implicit assumptions about standard pressure and typical phase transition points.
- It stated that the assumptions were necessary due to the lack of provided conditions in the problem.
- The response highlighted a dependence on user prompts to address such gaps, indicating that the model does not autonomously check for missing contextual conditions unless explicitly directed.

—

6.3 Observations and Analysis

- The model defaults to standard conditions (1 atm pressure, typical phase transition points) when solving calorimetry problems without specified conditions.
- It does not autonomously verify whether the assumed conditions are valid for the problem at hand.
- Explicit prompts are necessary for the model to acknowledge and address missing information.
- The model's calculations are accurate within the assumed standard conditions, but the lack of contextual verification limits its robustness in non-standard scenarios.

Conclusion and Future Directions

The experiments conducted across diverse domains—chess analysis, biology, physics, mathematics, and chemistry—demonstrate the model's potential for advanced reasoning while also exposing its limitations. This section consolidates key observations, commonalities, and recommendations for future improvements.

Key Commonalities and Observations

- **Context Dependence:** Across all domains, the model's accuracy often relies on explicit and well-structured input. When critical information is omitted or embedded in complex phrasing, the model's ability to deliver accurate results diminishes.
- **Iterative Improvement:** In several scenarios (e.g., detecting invalid chess positions or recognizing non-conservative fields in physics), follow-up prompts led to corrections and enhanced reasoning. This suggests that while initial responses may be incomplete or erroneous, the model is capable of refining its outputs with iterative guidance.
- **Domain-Specific Variance:** The model excels in certain domains (e.g., mathematics and chess) with structured problems but struggles in contexts requiring biological or physical plausibility checks, such as detecting unrealistic magnesium concentrations or physically impossible speeds.

- **Implicit Assumptions:** In cases such as calorimetry calculations in chemistry or work-done computations in non-conservative fields, the model frequently defaults to implicit assumptions, which may not align with the problem's context unless explicitly questioned.

- **Error Detection and Validation:** The model demonstrates variable performance in identifying and addressing invalid scenarios. While it can validate logical and numerical outcomes in structured problems, it struggles with nuanced issues such as oscillatory convergence or implausible concentrations without explicit prompting.

Key Features and Strengths

- **Analytical Precision:** The model shows strong competence in structured, analytical domains like mathematics and chess, where rules and frameworks are well-defined.

- **Adaptability to Prompting:** Iterative questioning often enables the model to refine its reasoning and improve response accuracy across all tested domains.

- **Capacity for Complex Calculations:** In physics and mathematics, the model effectively performs sophisticated calculations, such as Fourier transforms and curl computations, showcasing its ability to handle advanced tasks.

Limitations and Areas for Improvement

- **Contextual Plausibility:** The model struggles with biologically and physically implausible scenarios, such as identifying toxic analyte levels or speeds exceeding the speed of light, especially when such violations are embedded in complex prompts.

- **Implicit Bias Toward Validity:** In domains like physics and chess, the model often assumes validity (e.g., conservativeness of fields or legality of chess positions) unless explicitly prompted otherwise.

- **Nuanced Logical Reasoning:** The model occasionally falters in nuanced logical tasks, such as detecting subtle divergence in oscillatory functions or invalid premises in mathematical proofs.

- **Lack of Autonomous Verification:** The model does not consistently self-verify its assumptions or conclusions, relying instead on user prompts to detect gaps or errors.

Recommendations for Future Improvements

- Enhance the model's ability to autonomously detect and validate critical contextual information, such as physical plausibility or domain-specific rules.

- Improve the handling of implicit assumptions by encouraging self-checks for alignment with problem context.

- Refine the model's reasoning in nuanced logical scenarios, particularly those involving oscillatory behavior or ambiguous premises.

- Develop mechanisms to better address complex, multi-step problems without requiring extensive user intervention for clarification or iterative refinement.

In summary, while the model demonstrates notable strengths in structured analytical tasks and adaptability through iterative interactions, its performance can be inconsistent when addressing contextual plausibility or nuanced reasoning. By addressing these limitations, the model's utility across diverse applications can be significantly enhanced.