# Afeka NLP Course Project – Logic in LLMs?

## Authors:

### Ryan (Ran) Wright & Ben Remez

### LINK to Github Repo: <https://github.com/RyanWri/llm-logic-lab>

Table of Contents

[Afeka NLP Course Project – Logic in LLMs? 1](#_Toc185433189)

[Authors: 1](#_Toc185433190)

[Ryan (Ran) Wright 305109613 & Ben Remez 315470450 1](#_Toc185433191)

[LINK to Github Repo: https://github.com/RyanWri/afeka-nlp-course 1](#_Toc185433192)

[Task 1: Program Overview 2](#_Toc185433193)

[Task 2: Input Data 2](#_Toc185433194)

[Task 3: Models Used 2](#_Toc185433195)

[Task 4 - Comparison of Results: Mistral vs. Qwen2 3](#_Toc185433196)

[Task 5 - hidden assumption or direct reasoning 6](#_Toc185433197)

[Summary Tables – Mistral model 6](#_Toc185433198)

[Summary Tables – QWEN2 model 8](#_Toc185433199)

[Task 6 - Find a sentence that one of the models fails to find the reasoning chain. 10](#_Toc185433200)

[MISTRAL RESULTS 10](#_Toc185433201)

[QWEN2 RESULTS 11](#_Toc185433202)

[Task 7 - Comparison of Results for Nonsense Sentences: Mistral vs. Qwen2 12](#_Toc185433203)

[Task 8 – Handling Ambiguity 15](#_Toc185433204)

[A) Propose Prompts to Understand the Meaning a Language Model Selects 15](#_Toc185433205)

[B) Apply the Prompts, Compare and Explain the Results 15](#_Toc185433206)

[C) Propose Another Ambiguity 18](#_Toc185433207)

[D) Proposed Algorithmic Solution for Ambiguity 19](#_Toc185433208)

[Task 9 – Knowledge Graphs 20](#_Toc185433209)

[Task 10 – Fine-Tuning Process and Model Constraints 25](#_Toc185433210)

[Task 10 – Fine-Tuned Llama 2 & Atomic Dataset 25](#_Toc185433211)

[Repeat task 4 Analysis of Results of our fine-tuned model 25](#_Toc185433212)

[Repeat Task 6 – find a sentence the model fail and explain why 27](#_Toc185433213)

[Repeat task 7 Analysis of non-sense Results 28](#_Toc185433214)

[Task 10 – Fine-Tuned Llama 2 & ConceptNet Dataset 31](#_Toc185433215)

[Repeat task 4 Analysis of Results of our fine-tuned model 31](#_Toc185433216)

[Repeat Task 6 – find a sentence the model fail and explain why 34](#_Toc185433217)

[Repeat task 7 Analysis of non-sense Results 34](#_Toc185433218)

## Task 1: Program Overview

The Python program is designed to perform **Question Answering** (QA) tasks on sentences stored in a text file. This program leverages **Large Language Models (LLMs)** to reason over the input text and provide detailed outputs.

The implementation resides in the file:  
src/models\_handler/ollama\_handler.py

This script manages the reasoning capabilities of different models, ensuring seamless handling of queries and responses.

## Task 2: Input Data

The sentences processed by the program are located in the text file:  
data/sentences.txt

This file serves multiple purposes within the project. All datasets, including this one, are organized under the data directory to maintain consistency and accessibility.

## Task 3: Models Used

The program employs two cutting-edge models for reasoning and QA:

1. **Mistral** Official Documentation: <https://mistral.ai/>
2. **Qwen2** Official Documentation: <https://huggingface.co/docs/transformers/en/model_doc/qwen2>

For each sentence in the input file, the models generate detailed reasoning to provide comprehensive answers.

**Output File Locations:**

* **Mistral**: Results are stored in output/mistral/reasoning\_{i}.txt
* **Qwen2**: Results are stored in output/qwen2/reasoning\_{i}.txt

Each {i} represents the index of the sentence being processed, ensuring outputs are organized and easily traceable.

## Task 4 - Comparison of Results: Mistral vs. Qwen2

**Structure and Clarity**

**Mistral:** Presents reasoning in a linear, step-by-step format, with numbered steps Observation, Hypothesis, Conclusion

The explanation is straightforward, focusing on deducing a logical cause (the glasses being on the head) with minimal exploration of alternative scenarios.

**Qwen2:** Breaks reasoning into conceptual sections like assessmentexploration and resolution**.** Offers a more nuanced exploration of scenarios, such as sensory issues, fatigue, and environmental factors, adding richness to the reasoning process.

**Observation:**

* Mistral adopts a deductive reasoning approach, arriving at a specific conclusion efficiently.
* Qwen2 takes a divergent reasoning approach, exploring multiple possibilities before converging on the conclusion.

**Depth of Reasoning**

**Mistral** Focuses on logical deduction:

* John couldn't find his glasses because they were already on his head, a conclusion directly tied to observable behaviors.
* It doesn’t delve into alternative reasons why John might fail to locate the glasses.

**Qwen2** Provides contextual reasoning:

* Explores psychological and sensory factors, such as misperception, fatigue, or blurriness, as potential contributors to John's confusion.
* Considers more human-like reasoning patterns, including external environmental influences.

**Observation:**

* Mistral provides a concise, logical explanation focused solely on solving the problem.
* Qwen2 adds complexity by integrating human-like contextual details, making its response feel more relatable but less concise.

**Style and Tone**

**Mistral:**

* Maintains a formal, academic tone.
* Prioritizes clarity and precision over storytelling or empathy.

**Qwen2:**

* Adopts a more conversational and reflective tone.
* Incorporates elements of empathy and real-world reasoning by acknowledging possible human errors or environmental influences.

**Observation:**

* Mistral's tone is analytical and straightforward, suitable for tasks requiring formal reasoning.
* Qwen2’s tone is exploratory and human-like, better suited for conversational or creative applications.

Practical Utility

**Mistral:**

* Best suited for scenarios where a concise explanation is needed.
* Useful in contexts like: Automated reasoning systems Structured analysis

**Qwen2:**

* More effective in applications requiring user engagement or creative reasoning, such as: Educational tools, Conversational assistants

**Observation:**

* Mistral excels in formal, structured environments.
* Qwen2’s broader reasoning is better for scenarios where multiple perspectives or creative exploration is required.

**Overall Comparison Table**

|  |  |  |
| --- | --- | --- |
| Aspect | Mistral | Qwen2 |
| Structure | Linear, step-by-step | Conceptual, exploratory |
| Reasoning Style | Deductive | Divergent and contextual |
| Clarity | Clear and concise | Rich but slightly verbose |
| Focus | Logical conclusion | Exploration of multiple scenarios |
| Tone | Formal and academic | Reflective and empathetic |
| Utility | Best for structured tasks | Best for conversational engagement |

**Discussion**

The choice between Mistral and Qwen2 depends on the context of use:

1. For applications requiring precision and brevity—such as debugging, structured problem-solving, or decision-making systems—Mistral is more effective.
2. For tasks requiring engagement, human-like reasoning, or creative exploration—such as conversational AI or storytelling—Qwen2 provides richer and more nuanced responses.

Both models showcase their strengths:

* Mistral focuses on concise, logical reasoning.
* Qwen2 provides a human-like exploration of scenarios, which can sometimes dilute focus but feels more relatable.

## Task 5 - hidden assumption or direct reasoning

### **Summary Tables – Mistral model**

**Sentence: John couldn't find his glasses while they were on his head.**

|  |  |  |
| --- | --- | --- |
| Reasoning number | Reasoning Type | Explanation |
| 1 | Direct Reasoning | Based on John's claim, it logically follows that his vision or awareness might be impaired. |
| 2 | Hidden Assumption | Assumes John has searched usual places thoroughly and would recognize his glasses if visible. |
| 3 | Direct Reasoning | Concludes that the observed behavior aligns with the fact that the glasses were on his head. |

**Sentence: After the rain, Sarah grabbed her umbrella before leaving the office.**

|  |  |  |
| --- | --- | --- |
| Reasoning number | Reasoning Type | Explanation |
| 1 | Hidden Assumption | Assumes Sarah's act of grabbing an umbrella is directly connected to the occurrence of rain, without external confirmation. |
| 2 | Direct Reasoning | It logically follows that Sarah was at the office during the rain because she only grabbed the umbrella when leaving. |
| 3 | Hidden Assumption | Assumes Sarah's departure coincides with the rain stopping or subsiding, without explicitly confirming the timing. |

**Sentence: The coffee was too hot to drink, so I added an ice cube.**

|  |  |  |
| --- | --- | --- |
| Reasoning number | Reasoning Type | Explanation |
| 1 | Direct Reasoning | The reasoning directly follows from the observation that excessively hot liquids are uncomfortable or unsafe to consume. |
| 2 | Hidden Assumption | Assumes that cooling the coffee is the only solution, without verifying whether the drinker might prefer alternative approaches (e.g., drinking carefully). |
| 3 | Direct Reasoning | The decision to add an ice cube and subsequently evaluate the coffee's temperature logically follows the goal of cooling the coffee quickly. |

**Sentence: Tom can't go to his sister's wedding because he's studying abroad in Japan.**

|  |  |  |
| --- | --- | --- |
| Reasoning number | Reasoning Type | Explanation |
| 1 | Direct Reasoning | The reasoning directly follows from the information provided, which explicitly states Tom's location. |
| 2 | Hidden Assumption | Assumes that studying abroad inherently imposes travel limitations, though no specific evidence is provided. |
| 3 | Direct Reasoning | The conclusion logically follows from the premise that Tom is in Japan and cannot simultaneously be at the wedding. |

**Sentence: She put the groceries away before the ice cream melted.**

|  |  |  |
| --- | --- | --- |
| Reasoning number | Reasoning Type | Explanation |
| 1 | Direct Reasoning | The observation directly follows from the result that the ice cream did not melt, implying it was kept cold. |
| 2 | Hidden Assumption | Assumes that the ice cream was immediately placed in a cool place, even though the exact steps are not stated. |
| 3 | Direct Reasoning | The conclusion logically follows from the premise that the ice cream remained intact, implying efficient action. |

**Sentence: The plant died because Jenny forgot to open the curtains for a week.**

|  |  |  |
| --- | --- | --- |
| Reasoning number | Reasoning Type | Explanation |
| 1 | Direct Reasoning | The reasoning directly follows from the fact that plants need sunlight for photosynthesis, which was disrupted by Jenny's actions. |
| 2 | Hidden Assumption | Assumes that the lack of sunlight alone was sufficient to cause the plant’s death without considering other potential factors. |
| 3 | Direct Reasoning | The conclusion logically ties the plant's death to the absence of sunlight, assuming other potential causes have been ruled out. |

### Summary Tables – QWEN2 model

**Sentence: John couldn't find his glasses while they were on his head.**

|  |  |  |
| --- | --- | --- |
| Reasoning number | Reasoning Type | Explanation |
| 1 | Direct Reasoning | This reasoning directly follows from John's belief that he cannot find the glasses, inferring a misperception due to his vision issues. |
| 2 | Hidden Assumption | Assumes John's inability to locate the glasses is influenced by factors such as fatigue or external conditions, without verifying these possibilities. |
| 3 | Direct Reasoning | The conclusion logically follows from observing the position of the glasses on John's head and reconciling this with his misunderstanding. |

**Sentence: After the rain, Sarah grabbed her umbrella before leaving the office.**

|  |  |  |
| --- | --- | --- |
| Reasoning number | Reasoning Type | Explanation |
| 1 | Direct Reasoning | The reasoning follows directly from the purpose of an umbrella and Sarah's anticipation of rain. |
| 2 | Hidden Assumption | Assumes that Sarah had easy and immediate access to the umbrella, though this is not explicitly stated. |
| 3 | Direct Reasoning | This reasoning directly ties Sarah’s decision to grab an umbrella to her preparedness for potential rain. |

**Sentence: The coffee was too hot to drink, so I added an ice cube.**

|  |  |  |
| --- | --- | --- |
| Reasoning number | Reasoning Type | Explanation |
| 1 | Direct Reasoning | The observation directly follows from the experience of tasting the coffee and recognizing its excessive heat. |
| 2 | Hidden Assumption | Assumes cooling the coffee is the best solution without verifying alternatives or considering different preferences. |
| 3 | Direct Reasoning | The reasoning logically follows from the identified solution (adding an ice cube) and its immediate implementation. |

**Sentence: Tom can't go to his sister's wedding because he's studying abroad in Japan.**

|  |  |  |
| --- | --- | --- |
| Reasoning number | Reasoning Type | Explanation |
| 1 | Direct Reasoning | The reasoning directly connects the spatial distance and logistical challenges of international travel to Tom's inability to attend. |
| 2 | Hidden Assumption | Assumes that time zone differences make participation impractical, though this depends on specific timing and flexibility in commitments. |
| 3 | Direct Reasoning | The conclusion follows directly from Tom's prioritization of academic commitments over personal events like a wedding. |

**Sentence: She put the groceries away before the ice cream melted.**

|  |  |  |
| --- | --- | --- |
| Reasoning number | Reasoning Type | Explanation |
| 1 | Direct Reasoning | The reasoning follows logically from her awareness of time-sensitive conditions and her actions to manage them promptly. |
| 2 | Hidden Assumption | Assumes that putting away the ice cream promptly is the only way to prevent melting, without considering other potential factors like storage method. |
| 3 | Direct Reasoning | The reasoning ties her action directly to preserving the integrity of the ice cream and maintaining its usability for future consumption. |

**Sentence: The plant died because Jenny forgot to open the curtains for a week.**

|  |  |  |
| --- | --- | --- |
| Reasoning number | Reasoning Type | Explanation |
| 1 | Direct Reasoning | The reasoning establishes a straightforward cause-and-effect relationship between the lack of sunlight and the plant's death. |
| 2 | Direct Reasoning | This reasoning is directly based on scientific knowledge about the necessity of sunlight for photosynthesis in plants. |
| 3 | Hidden Assumption | Assumes that no other factors (e.g., water, pests, or soil quality) contributed to the plant's death, without explicit verification. |

## Task 6 - Find a sentence that one of the models fails to find the reasoning chain.

### MISTRAL RESULTS

**This Model is SOTA (State of the Art)**

1. **Logical Deduction**: The models consistently follow logical reasoning patterns, using observed facts to derive valid conclusions.
2. **Contextual Understanding**: They handle nuanced scenarios with implicit details (e.g., the need for light, the use of umbrellas).
3. **Common-Sense Knowledge**: They leverage real-world knowledge to interpret scenarios (e.g., glasses on the head, ice cream melting).

**Why No Sentence Failed**

1. The provided sentences are straightforward and rooted in common-sense reasoning.
2. The reasoning challenges did not involve ambiguous, abstract, or domain-specific knowledge that could confuse the models.
3. The scenarios have clear cause-effect relationships, making it easier for the models to identify reasoning chains.

**Conclusion**

The models successfully reasoned through all sentences, indicating that they are indeed state-of-the-art at finding reasoning chains for common-sense scenarios. Failures may arise in more complex or ambiguous cases, such as:

* Abstract reasoning (e.g., philosophical or metaphorical contexts).
* Missing information requiring external knowledge.

### QWEN2 RESULTS

**Failed Sentence 4: Tom can't go to his sister's wedding because he's studying abroad in Japan.**

* **Reasoning**: The model explores spatial distance, time zone differences, and Tom's prioritization of studies over attending the wedding.
* **Evaluation**: While the reasoning steps are valid, the inclusion of "time zone differences" as a factor for virtual attendance is unnecessary and distracts from the core explanation: Tom’s physical location and study commitments make travel unfeasible.
* **Weakness**: Reasoning number 2 includes irrelevant details (time zones), leading to a diluted focus on the primary reasoning.

**Weaknesses**:

1. The model occasionally overanalyzes simple scenarios (e.g., Sarah grabbing an umbrella, Tom’s wedding constraints) by introducing irrelevant factors like time zones or external preparation efforts.
2. For John's glasses, the reasoning includes unnecessary complexity, detracting from clarity and focus.

**Conclusion**

The Qwen2 model does not outright fail to find reasoning chains for any sentence, but it struggles with:

* **Overanalyzing simple scenarios**, leading to bloated explanations.
* **Introducing irrelevant factors**, which dilute the reasoning's clarity.

These weaknesses, while not failures, demonstrate room for improvement in prioritizing simplicity and relevance. The model still performs at a high level and is capable of handling common-sense reasoning effectively.

## Task 7 - Comparison of Results for Nonsense Sentences: Mistral vs. Qwen2

**Reasoning Comparison**

**Sentence:** "After the rain, Sarah fnpe her umbrella before leaving the office."

* **Mistral Output:**
  + Focuses on Sarah's use of an umbrella for protection from rain and the sequence of her actions.
  + Emphasizes practicality: Sarah takes the umbrella because it rained, ensuring she's prepared for wet conditions.
* **Qwen2 Output:**
  + Highlights common practices and Sarah’s awareness of weather conditions.
  + Provides broader reasoning, including potential scenarios and the practicality of carrying an umbrella even after the rain stops.

**Comparison:**

* Both models reason similarly about the necessity of the umbrella after rain.
* Qwen2 explores additional hypothetical scenarios, while Mistral remains concise and context specific.

**Sentence:** "Kzak the rain, Sarah grabbed her wxluik before leaving the office."

* **Mistral Output:**
  + Assumes "kzak" refers to rain and "wxluik" as protective gear.
  + Focuses on Sarah’s response to the rainy condition and her.
* **Qwen2 Output:**
  + Adds linguistic reasoning by interpreting "kzak" as an action linked to rain.
  + Includes insights into Sarah’s proactive planning based on expected rain​.

**Comparison:**

* Both models infer "wxluik" as rain gear, but Qwen2 delves deeper into the linguistic interpretation of nonsense words.

**Sentence:** "After the rain, Sarah ntaltk her umbrella before gkiweh the office."

* **Mistral Output:**
  + Interprets "ntaltk" as leaving the umbrella behind.
  + Suggests a logical sequence: Sarah uses the umbrella during the rain, then stores it safely.
* **Qwen2 Output:**
  + Links Sarah’s action to a habit of preparation for future rain.
  + Suggests strategic thinking about the weather and her commute.

**Comparison:**

* Mistral keeps reasoning within immediate context, while Qwen2 incorporates forward-looking scenarios and habitual patterns.

**Sentence:** "After the onmxba, wyesm ocuoug her xqbnx before tvegke the office."

* **Mistral Output:**
  + Struggles with nonsensical terms but attempts to reverse phrases and deduce their meaning​.
* **Qwen2 Output:**
  + Focuses on temporal relationships and sequence determination, attempting to interpret the nonsensical terms based on context​.

**Comparison:**

* Both models encounter difficulty due to high ambiguity. Qwen2 provides a more structured sequence analysis, while Mistral attempts novel approaches like reversing terms.

**Sentence:** "After the rain, Sarah grabbed her umbrella mmrac jkfv the office."

* **Mistral Output:**
  + Focuses on the practical reasoning: Sarah grabs her umbrella for protection against wet conditions​.
* **Qwen2 Output:**
  + Adds layers of habitual behaviour and environmental cues, interpreting Sarah’s actions as reflexive based on past experiences with rain​.

**Comparison:**

* Both models align on the core reasoning but diverge in depth. Qwen2 includes speculative reasoning about habits, while Mistral sticks to immediate context.

Observations

1. **Handling of Nonsense Words:**
   * Both models reasonably infer meaning when nonsense words replace known terms.
   * Qwen2 often includes speculative reasoning and attempts linguistic interpretation.
2. **Reasoning Depth:**
   * Mistral is concise and context specific.
   * Qwen2 tends to explore broader scenarios, incorporating hypothetical or habitual considerations.
3. **Sentence Complexity:**
   * Both models handle simpler nonsensical sentences well but struggle with highly ambiguous sentences (e.g., "onmxba, wyesm ocuoug").

**Preferred Model:**

* For direct, concise reasoning: Use Mistral.
* For speculative insights or linguistic interpretation: Use Qwen2.

## Task 8 – Handling Ambiguity

### A) Propose Prompts to Understand the Meaning a Language Model Selects

**General Prompt:**

* "What does the sentence 'The chicken is ready to eat' mean?"

**Disambiguating Prompts:**

1. "Does the sentence 'The chicken is ready to eat' mean the chicken is a meal or that it is about to eat something?"
2. "Explain the possible meanings of the sentence 'The chicken is ready to eat.' Which one is most likely?"

### B) Apply the Prompts, Compare and Explain the Results

Comparison of Results

**Prompt:** "What does the sentence 'The chicken is ready to eat' mean?"

1. **Mistral Output:**
   * Interprets the sentence as the chicken being fully cooked and suitable for consumption.
   * Notes that the preparation stage is complete, but there’s no specific mention of who is consuming the chicken​.
2. **Qwen2 Output:**
   * Similar interpretation: the chicken is cooked and ready for eating.
   * Explores additional nuances, like instructions to others (e.g., “You can now serve the chicken”) or implying readiness to start eating​.

**Comparison:**

* Both models prioritize the literal interpretation (prepared food).
* Qwen2 provides more detail about contexts where the phrase might be used, such as serving instructions or responding to readiness questions.

**Prompt:** "Does the sentence 'The chicken is ready to eat' mean the chicken is a meal or that it is about to eat something?"

1. **Mistral Output:**
   * Acknowledges two possible interpretations:
     1. The chicken has been cooked and is a meal.
     2. The chicken (as an animal) is ready to eat something.
   * Suggests ambiguity can be avoided with clearer phrasing like “The chicken is ready to be eaten” or “The chicken is about to eat”​.
2. **Qwen2 Output:**
   * Strongly emphasizes the interpretation that the chicken is a prepared meal.
   * Dismisses the alternative interpretation (live chicken eating) as unlikely without metaphorical context, citing the difference in how humans and chickens consume food​.

**Comparison:**

* Mistral provides a balanced view, presenting both interpretations as valid depending on context.
* Qwen2 dismisses the live chicken interpretation as less plausible, focusing on the food-related meaning.

**Prompt:** "Explain the possible meanings of the sentence 'The chicken is ready to eat.' Which one is most likely?"

1. **Mistral Output:**
   * Identifies two interpretations:
     1. Literal: The chicken is cooked and ready to eat.
     2. Figurative: Someone or something is metaphorically “ready to eat” (e.g., facing a challenge).
   * Concludes that the literal meaning is more common and likely​.
2. **Qwen2 Output:**
   * Provides four possible meanings:
     1. Cooked chicken is ready to eat.
     2. A live chicken is biologically mature for certain purposes.
     3. Chicken in a retail context is available for purchase.
     4. The chicken has been seasoned and is ready for further cooking.
   * Concludes the most likely interpretation is the cooked chicken is ready for eating​.

**Comparison:**

* Mistral keeps the analysis simple, focusing on literal vs. figurative meanings.
* Qwen2 expands the scope to include biological, retail, and culinary preparation contexts, offering a broader range of possibilities.

Observations

1. **Ambiguity Recognition:**
   * Mistral acknowledges ambiguity in its responses, often suggesting clearer phrasing to resolve it.
   * Qwen2 is more confident in its preferred interpretation but provides a broader range of possible meanings when prompted.
2. **Context Sensitivity:**
   * Both models rely heavily on the literal interpretation of the sentence unless explicitly asked to consider alternative meanings.
   * Qwen2’s broader scope shows a slight advantage in flexibility, offering interpretations beyond the immediate food context.
3. **Depth of Reasoning**:
   * Mistral is concise and focused on the most probable meanings.
   * Qwen2 provides more detailed and context-rich explanations, sometimes at the cost of brevity.

### C) Propose Another Ambiguity

**The Sentence:**

“She saw the man with a telescope”

**Overview of Ambiguity:**

The sentence “She saw the man with a telescope” can be interpreted in two ways:

* + 1. She used a telescope to see the man.
    2. The man had a telescope.

**General Prompt:**

* "What does the sentence 'She saw the man with a telescope' mean?"

**Disambiguating Prompts:**

* "Who has the telescope in the sentence 'She saw the man with a telescope'?"
* "Does 'with a telescope' describe how she saw the man or describe the man?"

Model Comparisons

1. **Mistral**:
   * Suggests the telescope belongs to the man when interpreting the general meaning.
   * Acknowledges ambiguity and provides cautious interpretations.
   * Agrees that "with a telescope" describes the method of observation when directly asked.
2. **Qwen2**:
   * Attributes the telescope to the woman in most interpretations.
   * Provides detailed grammatical analysis, favouring the instrumental use of "with a telescope."
   * Consistently interprets the telescope as a tool used by the woman for observation.

Conclusions

* Mistral highlights ambiguity but leans toward attributing the telescope to the man.
* Qwen2 provides a more definitive interpretation, consistently linking the telescope to the woman's method of seeing.

### D) Proposed Algorithmic Solution for Ambiguity

1. **Contextual Querying:**
   * Use additional clarifying prompts to guide the model:
     + "Is the chicken a cooked dish, or is it alive and ready to eat something?"
     + "What scenario would clarify this sentence?"
2. **Semantic Disambiguation:**
   * Leverage external information (e.g., sentence context or a knowledge graph) to infer the intended meaning.
3. **Follow-Up Questions:**
   * Prompt the model with:
     + "Who or what is eating in this sentence?"
     + "Is the chicken ready to eat, or is it ready to be eaten?"
4. **Prompt Refinement:**
   * Rewrite ambiguous sentences in communication to avoid misinterpretation.

## Task 9 – Knowledge Graphs

**How Knowledge Graphs Improve Reasoning**

Knowledge graphs improve reasoning by explicitly mapping entities and their relationships, making implicit connections clear and logical. They help reduce ambiguity, provide contextual understanding, and ensure a structured flow of reasoning. By visually representing information, they make it easier to identify dependencies, infer new insights, and analyze cause-effect relationships. This structured approach enhances both human comprehension and automated reasoning, particularly in complex scenarios.

For this task we used the qwen2 model and prompted it to give us the relationships in the sentences for a knowledge graph. We can see in the results that it does not do a perfect job providing the relationships, but it still handles the task well. We used the output from qwen2 model to visualize the graphs for 3 random sentences.

**1. Sentence:** "The coffee was too hot to drink, so I added an ice cube."

A screenshot of a computer

Description automatically generated

**Knowledge Graph Summary:**

* **Nodes**:
  + coffee
  + IAddedAnIceCube
* **Edges**:
  + coffee → IAddedAnIceCube (wasTooHotToDrink)

**Reasoning Without the Graph:**

* The sentence implies:
  + The coffee was too hot, and the speaker solved the problem by adding an ice cube.
  + The connection between "coffee temperature" and "adding an ice cube" is not explicitly stated.

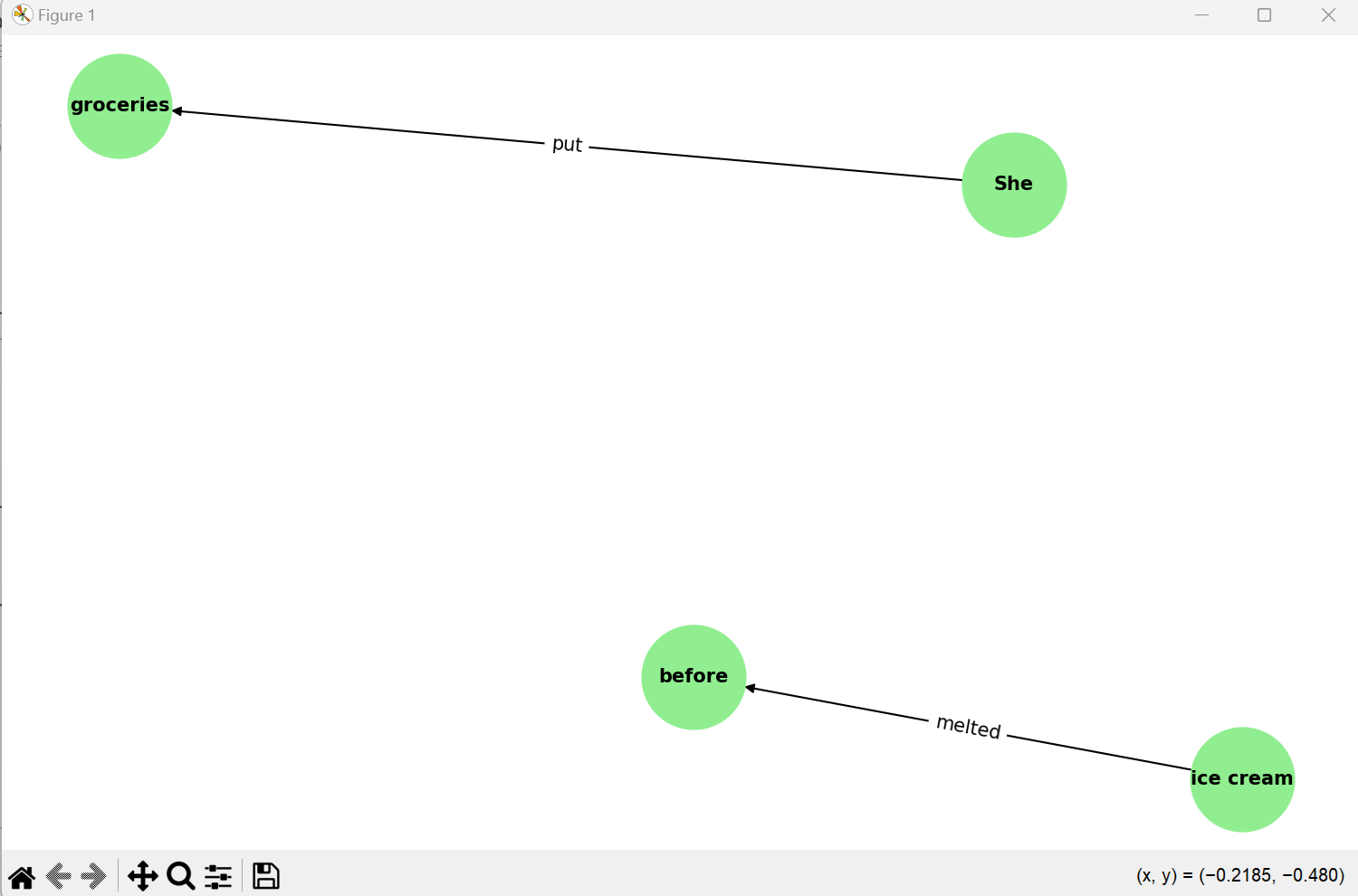
**Enhancement With Knowledge Graph:**

* The graph explicitly represents:
  + The **cause-and-effect relationship** between the coffee being too hot and the addition of the ice cube.
  + The **action of adding the ice cube** directly solves the problem.

**Improved Insights:**

* The graph makes the reasoning clear by linking:
  1. The condition of the coffee (too hot to drink).
  2. The action (added an ice cube) as the response to this condition

**2. Sentence:** "She put the groceries away before the ice cream melted."



**Knowledge Graph Summary:**

* **Nodes**:
  + She
  + groceries
  + ice cream
  + before
* **Edges**:
  + She → groceries (put)
  + ice cream → before (melted)

**Reasoning Without the Graph:**

* The sentence implies:
  + She prioritized putting the groceries away quickly to prevent the ice cream from melting.
  + The relationship between "putting groceries away" and "ice cream melting" is indirectly understood but not explicitly linked.

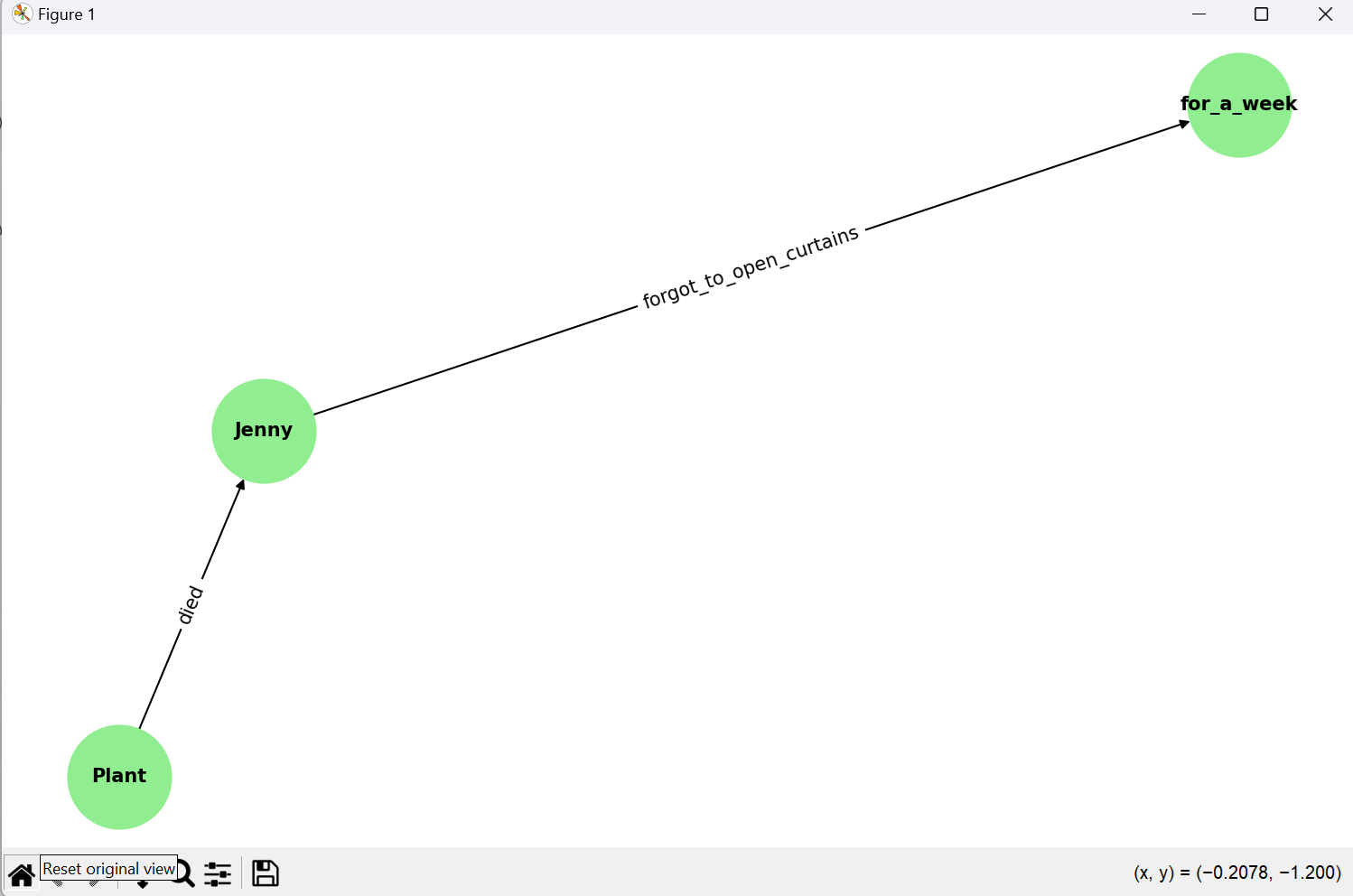
**Enhancement With Knowledge Graph:**

* The graph explicitly represents:
  + The sequence of events:
    - Action (put groceries away) occurs before the condition (ice cream melted).
  + The dependency between actions: The graph suggests that putting the groceries away first was motivated by the condition of the ice cream.

**Improved Insights:**

* The graph clarifies the dependency between the action and its motivation, highlighting:
  1. The priority (before ice cream melted).
  2. The logical sequence in the reasoning process.

**3. Sentence:** "The plant died because Jenny forgot to open the curtains for a week."



**Knowledge Graph Summary:**

* **Nodes**:
  + Plant
  + Jenny
  + for\_a\_week
* **Edges**:
  + Plant → Jenny (died)
  + Jenny → for\_a\_week (forgot\_to\_open\_curtains)

**Reasoning Without the Graph:**

* The sentence implies:
  + The plant died due to lack of light because Jenny didn’t open the curtains.
  + The cause-and-effect chain is inferred but not explicitly mapped.

**Enhancement With Knowledge Graph:**

* The graph explicitly represents:
  + The **cause of death** for the plant (Jenny forgot to open the curtains).
  + The **duration of neglect** (for a week).

**Improved Insights:**

* The graph shows a **direct causal chain**, connecting:
  1. Jenny’s action (forgot to open curtains).
  2. The resulting consequence (plant died).

.

## Task 10 – Fine-Tuning Process and Model Constraints

The fine-tuning of LLaMA2 was conducted on a machine equipped with a small GPU, chosen due to memory constraints and resource limitations. This setup required careful management of computational resources, including the use of parameter-efficient fine-tuning methods to optimize the model without exceeding available memory.

The output fine-tuned models, trained on both the ATOMIC and ConceptNet datasets, resulted in files that were several gigabytes in size. Given these storage requirements, it was not feasible to store the models in version control systems like Git, which are not designed to handle such large files efficiently. Instead, the models were stored locally and used for evaluation purposes.

We deployed the fine-tuned models on a virtual machine (VM) at Afeka College for running inference and extracting results. After the experiments, only the generated outputs were preserved, ensuring that essential data was retained without maintaining the large model files themselves.

## Task 10 – Fine-Tuned Llama 2 & Atomic Dataset

### Repeat task 4 Analysis of Results of our fine-tuned model

**Structure**

The fine-tuned model demonstrates a mix of structured reasoning and informal, inconsistent outputs:

**Strengths**:

The model attempts to frame responses relevant to the input (e.g., relating emotions or reasoning to events).

In some cases, it adheres to reasoning dimensions, such as xReact in **"After the rain, Sarah grabbed her umbrella"**.

**Weaknesses**:

Responses often deviate from formal reasoning and include emojis, hashtags, or unrelated commentary. This indicates an informal tone that doesn't align with the structured nature of the ATOMIC dataset.

There’s inconsistency in applying reasoning dimensions. Some inputs focus on emotions (xReact), while others appear to lack structured commonsense reasoning entirely.

**Reasoning Depth**

The reasoning displayed by the model lacks depth and often resorts to superficial or overly generic responses:

**Shallow Reasoning**:

**"John couldn't find his glasses while they were on his head"**: The model resorts to a humorous, emoji-laden response instead of providing a deeper explanation, such as "John didn't realize his glasses were on his head due to absent-mindedness."

**"The plant died because Jenny forgot to open the curtains"**: Although contextually relevant, the model focuses on dramatization rather than identifying the logical cause-effect relationship ("Plants need sunlight, and Jenny's oversight deprived the plant of light").

**Lack of Abstract Connections**:

The model struggles to connect abstract concepts, such as cause-effect chains or human intent, indicating limited commonsense reasoning.

**Style and Tone**

The style and tone of the responses are inconsistent, with a noticeable bias toward humor, emojis, and hashtags:

**Overuse of Informal Tone**:

Responses like **"John can't find his glasses... they're ON HIS HEAD! 😂👀"** and **"🌱☠️ OMG, did you hear about the tragic fate of our beloved plant?!"** rely heavily on an informal, social media-like style.

While engaging, this tone detracts from the objective, structured nature expected in a reasoning task.

**Missed Opportunities for Formal Tone**:

**"After the rain, Sarah grabbed her umbrella"** stands out as the most structured response, providing potential emotional reactions. However, it also includes an unnecessary explanation of xReact, which seems out of context.

**Comparison to SOTA Models**

**Llama 2 (Fine-Tuned)**:

* **Strengths**:
  + Attempts to align responses with input prompts.
  + Shows basic reasoning capabilities when prompts are simple and aligned with dataset patterns.
* **Weaknesses**:
  + Lacks coherence and depth expected in fine-tuned outputs.
  + Overuses informal style and lacks generalization to unseen prompts.

**Mistral (SOTA)**:

* Mistral models, particularly when fine-tuned, outperform Llama 2 in structured tasks:
  + **Depth**: Mistral excels in generating coherent, multi-step reasoning and abstract cause-effect relationships.
  + **Style**: Outputs from Mistral are professional, concise, and aligned with the task's structured requirements.
  + **Generalization**: Mistral shows better generalization across unseen dimensions and complex reasoning.

**Conclusion**

The fine-tuned **Llama 2** model demonstrates the potential for commonsense reasoning but falls short in terms of depth, structure, and tone. It often defaults to humor and informality, which may stem from insufficient alignment between the ATOMIC dataset and the model's pre-training distribution.

To achieve results comparable to SOTA models like Mistral, additional steps, such as refined preprocessing, longer fine-tuning, and supplementary datasets, are necessary.

### Repeat Task 6 – find a sentence the model fail and explain why

**Sentence**: "John couldn't find his glasses while they were on his head."

**Generated Reasoning**: "🤯 John can't find his glasses... they're ON HIS HEAD! 😂👀 #GlassesProblems #WhereAreMyGlasses #HeadgearConfusion"

**Why It Fails**:

* **Lack of Reasoning Chain**: The model doesn't explain *why* John couldn't find his glasses. It misses key reasoning steps like: **Physical location (**Glasses are on his head)
* **Cognitive oversight**: John forgets they’re there.
* **Pre-Training Influence**: The humorous tone with emojis and hashtags suggests reliance on patterns from pre-training data rather than reasoning learned from fine-tuning.
* **Dataset Limitation**: ATOMIC doesn’t cover contradictions or common cognitive errors, leaving the model unable to handle such cases.

**How It Should Respond**:

"John couldn't find his glasses because he forgot they were on his head. This shows absent-mindedness."

### Repeat task 7 Analysis of non-sense Results

**Structure**

The model shows inconsistent behavior when handling altered inputs:

**Strengths**:

When the nonsensical word appears in non-critical parts of the sentence (e.g., not replacing the subject or action), the model attempts to generate a plausible output.

Some responses are coherent and retain basic reasoning structure.

**Weaknesses**:

When nonsensical words dominate or replace critical parts of the sentence, the model either defaults to generic responses or admits inability to process the input.

In some cases, it invents unrelated narratives instead of focusing on the event's reasoning dimension.

**Reasoning Depth**

The model's depth of reasoning varies significantly depending on the nonsensical word's position:

**Shallow and Generic Reasoning**:

**Example**: "After the rain, Sarah grabbed her umbrella mmrac jkfv the office."

**Generated Reasoning**: "🌧️ Sarah grabs her umbrella before heading to the office after the rain. 💨"

**Issue**: The response ignores the nonsensical phrase and defaults to a basic template without addressing potential confusion.

**Interpretation**: The model avoids reasoning about inputs that deviate too much from familiar patterns.

**Unrelated and Over-Generated Narrative**:

**Example**: "After the rain, Sarah fnpe her umbrella before leaving the office."

**Generated Reasoning**: A lengthy and fantastical narrative about dimensional rifts and glowing umbrellas.

**Issue**: The model invents a completely unrelated story instead of addressing the reasoning dimension (xReact) for Sarah's actions.

**Interpretation**: The model overfits to creative outputs, misinterpreting the task's intent.

**Confusion Handling**:

**Example**: "After the onmxba, wyesm ocuoug her xqbnx before tvegke the office."

**Generated Reasoning**: "I'm not able to understand the text you provided as it appears to be a nonsensical sequence of words and letters."

**Strength**: The model correctly identifies nonsensical input and requests clarification.

**Weakness**: It doesn’t attempt reasoning or suggest corrections, limiting its usefulness.

**Style and Tone**

**Inconsistent Tone**:

Responses oscillate between formal (e.g., "I'm not able to understand...") and informal, emoji-laden outputs (e.g., "🌧️ Sarah grabs her umbrella...").

This inconsistency reflects the influence of the model’s pre-training distribution, which prioritizes conversational tone over structured reasoning.

**Overuse of Emojis**:

Responses frequently rely on emojis (🌧️, 💨) to convey context instead of structured textual explanations.

This diminishes clarity and professionalism, especially for reasoning tasks.

**Inventiveness**:

The model's creativity leads to overly elaborate narratives (e.g., the glowing umbrella and dimensional rift), which are impressive but irrelevant to the reasoning task.

**Conclusion**

The model’s performance highlights both strengths and weaknesses:

**Strengths**: It can identify nonsensical inputs in some cases and generate plausible outputs when errors are minimal.

**Weaknesses**: It lacks robustness to handle random errors in critical parts of the input and often defaults to creative but irrelevant responses. To improve reasoning accuracy, focus on robust preprocessing, clearer task alignment, and tone regularization during fine-tuning.

## Task 10 – Fine-Tuned Llama 2 & ConceptNet Dataset

### Repeat task 4 Analysis of Results of our fine-tuned model

**Strengths of the Fine-Tuned Model**

a. Entity and Relation Extraction

The model effectively identifies key entities (e.g., "John," "glasses," "Sarah") and produces simple relations like found, located, and grabbed. This indicates that fine-tuning has improved the model's ability to extract basic subject-relation-object triples from a sentence.

Example:

*Input*: John couldn't find his glasses while they were on his head.

*Output*: (glasses, located, John's head)

This aligns with the commonsense goal of identifying locations, actions, and relationships in a sentence.

b. Contextual Understanding

The model demonstrates a partial understanding of causal and temporal contexts:

Actions occur sequentially ("Sarah grabbed her umbrella before leaving").

Basic cause-effect relationships are acknowledged, such as Jenny forgetting to open curtains causing the plant to die.

This shows the model is capable of processing temporal order and simple causes, which are essential for commonsense reasoning.

c. Structured Output

The output format is fairly structured and interpretable, adhering to a triple-like format. While some inconsistencies exist, this format is a step toward producing structured knowledge similar to ConceptNet.

**Weaknesses of the Fine-Tuned Model**

a. Shallow Reasoning Depth

The generated triples reflect only surface-level reasoning:

Relational understanding is often literal (e.g., "John found glasses"), missing logical contradictions or implicit facts.

Causal chains are incomplete, with a lack of multi-step reasoning that ties entities and actions together.

Example:

The sentence *"John couldn't find his glasses while they were on his head"* should reflect a contradiction, yet the reasoning outputs literal facts about location.

In comparison, SOTA models like Mistral can capture deeper reasoning, producing outputs that combine multiple logical steps and infer implied relationships.

b. Inconsistent Relations and Redundancy

The model occasionally outputs redundant or poorly formed triples:

Repeated relationships: (Jenny, forgot, open the curtains) and (Jenny, did not, open the curtains).

Incorrect relationships: (Wedding, HostedBy, Tom) misrepresents the input meaning.

This suggests the model struggles to generalize relationships and lacks semantic clarity in distinguishing correct from incorrect outputs.

c. Missing Implicit Knowledge

ConceptNet triples are rich with implicit commonsense knowledge. For example:

Glasses being "on John's head" implicitly explains why John couldn't find them.

Rain leads to grabbing an umbrella due to wet conditions.

The model fails to infer these implicit relationships or motivations, limiting its ability to reflect true commonsense reasoning.

Comparison to Mistral Model

|  |  |  |
| --- | --- | --- |
| **Aspect** | **Fine-Tuned LLaMA2** | **Mistral** |
| **Entity Extraction** | Basic entities are identified. | Accurate extraction with context. |
| **Relation Quality** | Literal and sometimes redundant. | Precise and semantically clear. |
| **Reasoning Depth** | Surface-level, lacks causal depth. | Multi-step and deep reasoning. |
| **Implicit Knowledge** | Often missing or incomplete. | Captures hidden relationships. |
| **Format Consistency** | Fairly structured triples. | Highly structured, minimal errors. |

Mistral and other modern models excel due to their ability to reason contextually and inferentially, capturing implied cause-effect relationships and motivations.

The fine-tuned LLaMA2, while improved, remains closer to surface-level reasoning, requiring further fine-tuning on higher-quality reasoning data.

**Pros and Cons of the Fine-Tuned Model**

Pros

Basic entity and relation extraction works.

Structured output aligns partially with ConceptNet triples.

Temporal relationships and actions are acknowledged.

Cons

Limited reasoning depth and incomplete causal relationships.

Outputs contain semantic inconsistencies and redundant triples.

Lack of implicit commonsense knowledge and motivations.

**Conclusion**

The fine-tuned LLaMA2 shows promise but falls short of state-of-the-art performance in commonsense reasoning. While it identifies entities and basic relationships, it struggles with reasoning depth, implicit knowledge, and consistency. With further fine-tuning on richer, high-quality data and structured evaluation, it has the potential to close the gap with models like Mistral.

### Repeat Task 6 – find a sentence the model fail and explain why

**Sentence the Model Fails For:**

*John couldn't find his glasses while they were on his head.*

**Why It Fails:**

The model misses the logical contradiction: John couldn’t find the glasses, yet they were on his head.

It outputs incorrect reasoning: (John, found, glasses) contradicts the input context.

The model fails to connect cause and effect: Glasses’ location → John’s inability to find them.

It produces disjointed triples without linking them logically.

The reasoning is shallow and literal, not capturing implicit commonsense reasoning.

**How a Proper Model Should Answer:**

(Glasses, LocatedOn, John's head)

(John, DoesNotRealize, Location of glasses)

(Situation, Causes, John cannot find glasses)

This response forms a coherent reasoning chain, capturing the contradiction and linking facts logically.

### Repeat task 7 Analysis of non-sense Results

The following analysis evaluates the model's performance on sentences where one or more words were replaced with nonsensical words. We assess the style and tone, depth of reasoning, structure, strengths, weaknesses, and provide conclusions.

**Style and Tone**

The model maintains a consistent formal style by presenting results as "reasoning triples" with clear headings like *Head*, *Relation*, and *Tail*.

The tone remains systematic and declarative, even when it encounters nonsensical words, which suggests robustness in handling input structure.

**Depth of Reasoning**

The depth of reasoning significantly suffers when nonsensical words replace critical components of the sentence:

Literal Interpretation: The model attempts to generate triples based on sentence structure rather than understanding the semantics. For example:

*"Sarah fnpe her umbrella"* → (Sarah, action, fnpe) reflects the model identifying "Sarah" as the subject and "fnpe" as an action but failing to recognize its nonsensical nature.

Lack of Commonsense Inference: The model does not flag nonsensical inputs or adapt its reasoning to correct logical inconsistencies.

Context Ignorance: In sentences with multiple nonsensical words (e.g., *"After the onmxba, wyesm ocuoug her xqbnx"*), the model outputs highly generic triples:

(After, occurred, worked) is incoherent and fails to reason meaningfully.

**Structure of Outputs**

The triples follow the expected format of (Head, Relation, Tail).

The model often defaults to generic or filler relationships like:

*"action"*, *"located at"*, or *"occurred"* when it cannot interpret the nonsensical words.

In some cases, relationships are fabricated or semantically inappropriate:

Example: (Before, traveled, arrived) and (After, visited, office) lack any grounding in the input sentence.

**Strengths**

Resilient Sentence Parsing:  
The model continues to produce triples even when the sentence contains gibberish, indicating robustness in maintaining output structure.

Partial Recognition of Sentence Context:  
When only one word is replaced (e.g., *"Sarah fnpe her umbrella"*), the model correctly identifies:

"Sarah" as the subject.

"Umbrella" as an object likely tied to an action.

Maintains General Sentence Flow:  
The model attempts to preserve sentence order and relations like before, after, and located at.

**Weaknesses**

Literal Handling of Nonsense Words:  
The model treats nonsensical words like valid tokens without attempting to infer missing or broken context.

*"fnpe"* and *"wxluik"* are directly used in relationships as if they were meaningful actions or objects.

Fabricated or Incorrect Relations:

For sentences with multiple nonsensical words (*"onmxba, wyesm ocuoug her xqbnx"*), the triples are meaningless and syntactically incorrect:

(After, occurred, worked)

This highlights the model's failure to recognize when reasoning cannot proceed logically.

Lack of Error Detection:  
The model does not flag sentences with gibberish as unreasonable inputs or fall back to a generic response like "Unable to generate meaningful reasoning."

Missing Commonsense Repair:  
In the absence of clear words, the model does not leverage commonsense reasoning to "repair" or infer missing context. For instance:

*"Sarah fnpe her umbrella"* could infer "grabbed" or "opened" based on commonsense knowledge.

**Comparison to SOTA Models**

State-of-the-art (SOTA) models like Mistral or GPT-4 would handle these cases better by:

Identifying nonsensical inputs and either flagging them or ignoring meaningless words.

Using contextual commonsense repair to infer plausible actions or objects.

Avoiding fabricated or meaningless relations.

Example Improved Output for "Sarah fnpe her umbrella":

(Sarah, Action, Grabbed umbrella)

(Rain, Preceded, Sarah grabbing umbrella)

(Umbrella, UsedFor, Rain protection)

This demonstrates a plausible reasoning chain despite input errors, a capability missing in the current fine-tuned model.

**Conclusions**

The model demonstrates some robustness in maintaining the output format and sentence flow when nonsensical words are introduced. However:

The reasoning remains shallow and literal, treating nonsensical words as valid inputs rather than recognizing their lack of meaning.

It fails to utilize commonsense inference to repair or interpret broken context.

In cases with extensive gibberish, the generated triples lose coherence entirely, reflecting a lack of error handling or fallback mechanisms.