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

## ****Section 4 Analysis of Results****

**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.

## Section 6

**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."

## ****Section 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.