# ****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.

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

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

**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.

### Section 6

**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.

## ****Section 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.