Problem Definition:

Our task is to fine-tune a T5 model to answer "why" questions based on a narrative. We do this by injecting additional commonsense context to our dataset using a larger model (Gemini).

Example Usage:

Input: {'narrative': "Cam ordered a pizza and discovered it wasn't pre-sliced.", 'question': 'Why did Cam order a pizza?'}

Output: "Cam was hungry."

Importance:

Understanding "why" requires reasoning beyond surface text, integrating context, and commonsense knowledge. Solving this problem is important as it advances models' ability to perform causal reasoning, which is critical for applications in education, conversational AI, and automated decision-making systems, etc.

Motivation:

This problem is typically solved by fine-tuning standard text generation models (e.g., T5, GPT) for question answering. These approaches focus on extracting answers directly from text without explicit reasoning or integrating external knowledge.

Gaps and Challenges:

Lack of reasoning: Many models fail to go beyond surface-level text understanding to infer causal relationships.

Contextual knowledge: Models struggle to incorporate external commonsense or domain-specific knowledge to enhance reasoning.

Limited evaluation metrics: Existing evaluations (e.g., BLEU, ROUGE) don’t fully capture reasoning quality or relevance.

Our approach bridges these gaps by integrating commonsense context into training and leveraging metrics like BLEURT, ROUGE, and Exact Match to measure reasoning capabilities.

Relevant Work:

"UnifiedQA: Crossing Format Boundaries with a Single QA System" (Khashabi et al., 2020) highlights the importance of diverse input formats.

"COMET: Commonsense Transformers" (Bosselut et al., 2019) emphasizes external knowledge integration for reasoning.

Our Main Ideas:

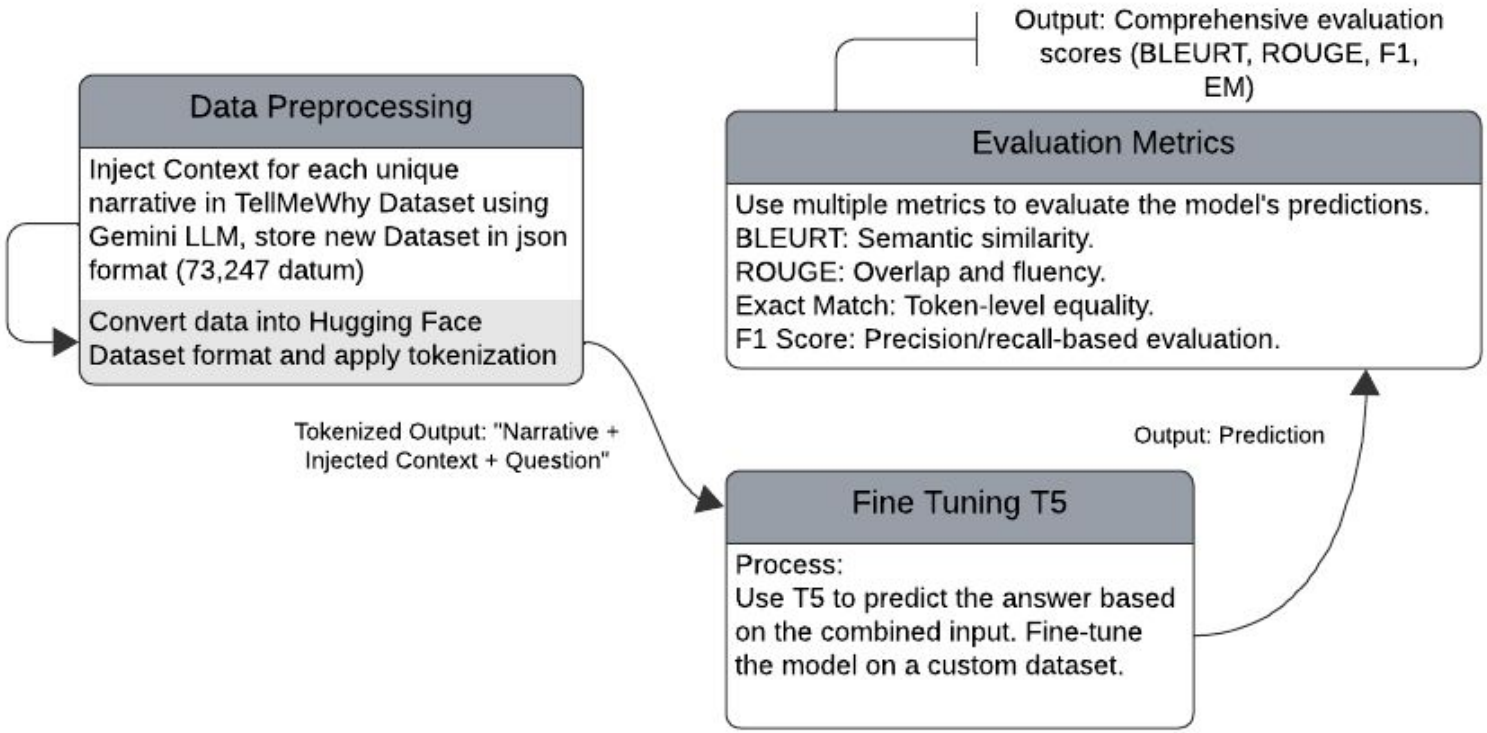
Context Injection for Enhanced Reasoning: We augment narrative-based inputs with external commonsense context using Gemini to improve the model's ability to infer causal relationships and answer "why" questions. This directly addresses the gap of insufficient reasoning capabilities by providing the model with additional explanatory knowledge.

Fine-Tuning T5 for Contextual Question Answering: Utilizing T5's pretrained generative abilities, we can produce answers that incorporate both the narrative and the injected context. This helps the model generate more accurate and context-aware answers, mitigating the limitations of extracting answers directly from text.

Improved Evaluation with BLEURT, ROUGE, and F1: We introduce evaluation metrics that assess semantic similarity, reasoning, and overall text quality. This provides us a more nuanced understanding of the model's performance beyond surface-level token matching.

These ideas address the gaps by implementing **reasoning enhancement** through injecting context, **generalization** by fine-tuning a T5 model on enriched inputs to improves its base capability, and **comprehensive evaluation** by using BLEURT and ROUGE to evaluate the quality and reasoning capability of generated answers, ensuring meaningful improvements over traditional approaches.

Method Details:



Evaluation setup:

Our Dataset–TellMeWhy–is a large-scale crowdsourced dataset made up of more than 30k questions and free-form answers concerning why characters in short narratives perform the actions described. We use this to create a custom dataset of context injected questions based on each datum.

Dataset: 73,247 samples of narratives, questions, injected commonsense context, and answer targets.

Some Dataset Statistics: {Input Length: Min=55, Max=256, Mean=131.13}, {Target Length: Min=3, Max=71, Mean=11.07}, {Most Frequent Input Length: 130 tokens}, {Most Frequent Target Length: 9 tokens}

Systems Compared: Baseline T5-small fine-tuned on the no context dataset *and* T5 model trained with additional context injection.

Evaluation Measures:

- BLEURT: Semantic similarity with a focus on commonsense relevance.
- ROUGE-L F1: Evaluates sequence-level overlap via longest common subsequence.
- Exact Match (EM): Percentage of predictions that match the reference exactly.
- F1 Score: Word-level harmonic mean of precision and recall.
- Loss Trends: Training and validation loss across epochs.

Goal: To assess the effectiveness of commonsense-enhanced training on question-answering performance through context injection.

Key Results: (Trained with 10,000 data examples from Stonybrook TellMeWhy dataset, 15% reserved for evaluation)

Fine Tuning a T5 with no context

Epoch	Train Loss	Validation Loss	BLEURT Score
1	0.170300	0.141289	-0.905466
2	0.148500	0.138184	-0.881379
3	0.142700	0.137338	-0.868918
4	0.139200	0.136923	-0.861303

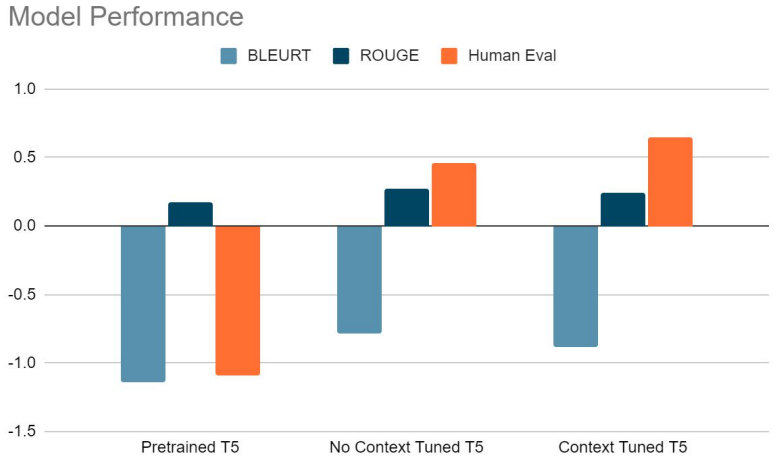
Fine Tuning a T5 with context added

Epoch	Train Loss	Validation Loss	BLEURT Score
1	0.166000	0.150395	-0.983019
2	0.156400	0.146696	-0.958194
3	0.147800	0.144871	-0.948719
4	0.146500	0.144614	-0.944824

BLEURT Scores for non-finetuned model: -1.1427

Average Human evaluated score on 100 questions (Scored from [-2,2]):

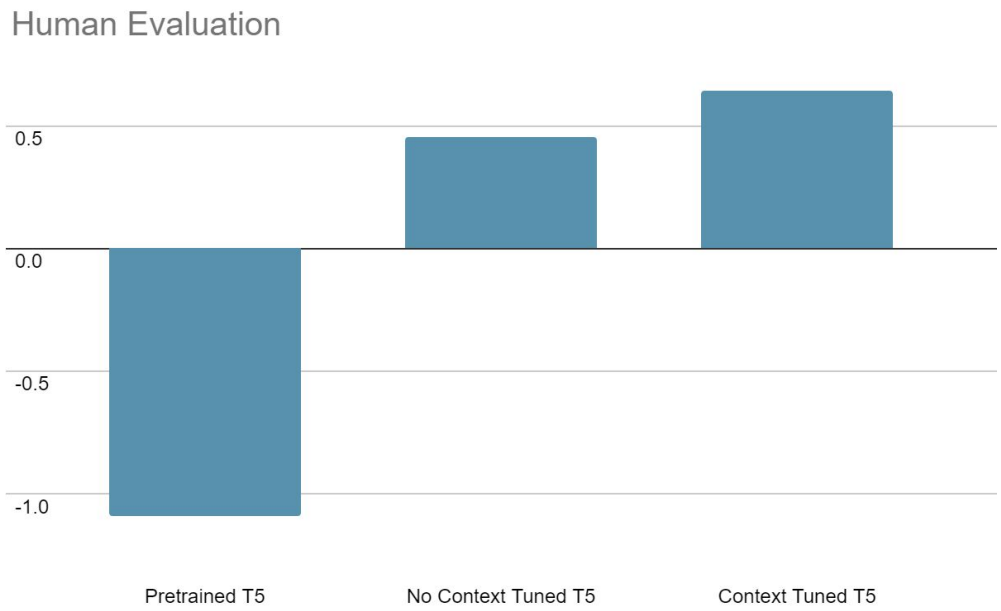
Raw pre-trained T5: **-1.090**
Trained with no context: **0.457**
Trained with context: **0.643**



Analysis:

We analyzed results using BLEURT, ROUGE-L, Exact Match, F1 Score, and 100 human evaluations. Automated metrics assessed semantic and token alignment, while human evaluations highlighted reasoning quality and contextual understanding.

Our system works well during human evaluation when the target value did not properly answer the question being asked and contextual understanding of the situation was needed. It however addressed nouns improperly or would reference the wrong section of text as answers, probably due to limited training sizes and exposure. The quality of the context is also worth mentioning, as some things would not be mentioned by our prompt or injection and thus could not be learned by the model in those instances, it is incredibly difficult to manage injecting context for all commonsense or meaningful knowledge pertaining to even a small text.



Conclusions:

Impact of Context Injection:

Adding context improved the model's performance slightly but increased token length, leading to limitations in sequence processing due to model constraints. More training data would be required to maximize the effectiveness of context injection. A larger model with more parameters would better help capture the context needed to answer the why questions.

Evaluation Metrics Insights:

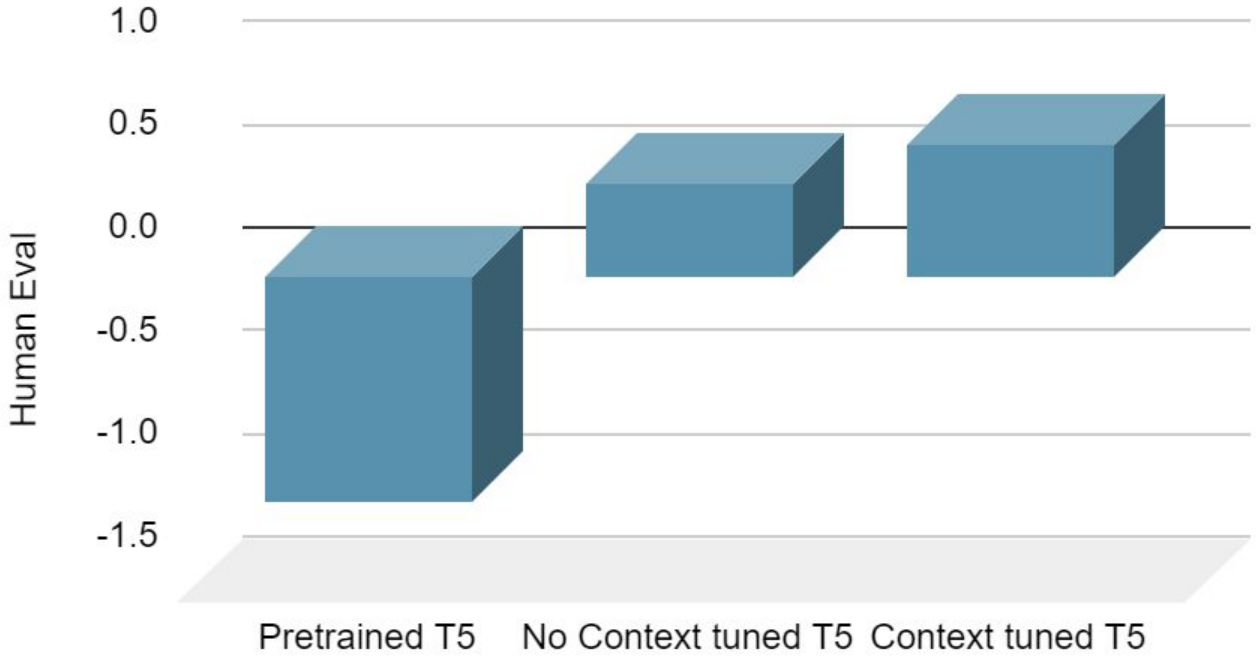
Human evaluations indicated the context-trained model produced reasonably better results. BLEURT scores, however, did not align with human evaluations, highlighting the limitations of BLEURT for this task. Metrics like ROUGE-L and BLEURT may adequately capture nuanced improvements introduced by context.

Challenges with Long-Sequence Models:

Token limitations of models like T5 prevented the model from fully leveraging long-context scenarios.Increasing sequence length would require more resources and likely larger models (e.g., T5-large or LongT5).

Future work would involve collecting larger datasets to enhance the effectiveness of context injection, exploring long-sequence models like LongT5, and refining evaluation with additional metrics like METEOR or BERTScore to better align with human judgment. Adaptive context injection strategies and domain-specific fine-tuning could further optimize token usage and improve reasoning in specific applications.

Human eval on no context



Human eval on no context