Fairytale Reading Comprehension

Dylan Sivori, Danielle Yoseloff
{dylan.sivori , dyoseloff} @berkeley.edu
August 2025

0 Abstract

In this paper, we explore narrative reading comprehension using the FairyTaleQA dataset. We focus on the performance gap between answering explicit (extractive) and implicit (abstractive) questions. To address this, we propose a pseudo Mixture of Experts (MoE) architecture, where smaller specialized models handle each question type. We experiment with fine-tuning Flan-T5 and QWEN-2-Instruct models as well as different prompting techniques like Chain of Thought (CoT) and evaluate them using ROUGE-L, BLEU, BLEURT, and our own qualitative analysis. Our results show that specialized fine-tuning outperforms a unified model, particularly in eliminating generic, low-quality answers. While metrics favor the Flan-T5 mixture, QWEN often produces more thoughtful, conversational responses, highlighting the need for both quantitative and qualitative evaluation in QA tasks.

1 Introduction

Reading comprehension is one of the most important skills that humans learn in their lifetime. It connects us to the rest of the world by allowing us to engage with others' thoughts, reflect on what we've read to form our own opinions, and learn from others' experiences to broaden our perspectives through understanding. Given this is such an important skill, this begs the question, how can we develop and evaluate one's level of reading comprehension? One of the most common practices around the world is to provide students with a passage and have them answer questions about the text to gauge their understanding.

There is currently a clear gap in question answering (QA) capability between humans and models when it comes to reading comprehension (Xu et al., ACL 2022). This gap is exemplified by the different types of questions that one may encounter. There are explicit questions that are more extractive and implicit questions that require some reasoning about the question given some context. The problem we try to solve is that these different question types have different requirements of a language model to answer them properly. This is particularly important because as humans begin to interact and rely more on language models, we need to make sure that the answers models return are accurate and reliable. Other than just the academic pursuit of improving capabilities in QA, we see this area of study having a number of educational applications. For example, teachers could use our QA model as a means of benchmarking student performance. They could see where students are struggling compared to the model output so they know where the students' pain points are and where they should focus their instruction. One other application we see is using our model as a way for teachers to gauge their question difficulty and structure their exams in a way to test students for both comprehension and reasoning abilities.

In our experiments, we set out to improve narrative question answering using the FairytaleQA dataset (Xu et al., ACL 2022). In their paper, they discuss the creation of their dataset and how it exposes the current shortcomings of different models and how they perform on different question types and narrative elements. In their experiments, they report on the ROUGE-L scores for different pre-trained models as well as fine-tuned models on the FairyTaleQA dataset. They also report on performance using the more well known NarrativeQA dataset (Kočiský et al., TACL 2018). We take a novel approach to this task and propose a pseudo MoE (Cai et al., 2025) to improve the quality of question answers. In other words, since our dataset is annotated to have two different question types, explicit and implicit, we propose to have smaller specialized models handle their respective question types so they can become experts at that task. This couldn't truly be solved without our approach because the way you answer these two question types is fundamentally different. We also propose that we should be evaluating model performance with more than just ROUGE-L scores. In addition to ROUGE-L, we will be reporting on BLEU and BLEURT scores to measure more faithfulness and semantic equivalence rather than just n-gram overlap. This will be especially important for implicit questions that require more reasoning and explanation versus extraction. In the following experiments we will explore different model architectures for different question types, as well as different prompting techniques in an attempt to enhance model performance.

2 Background

The challenge for models to perform reading comprehension is introduced to us by the creation of the NarrativeQA dataset (Kočiský et al., TACL 2018). Conventional methods for question answering tasks are reliant on information retrieval based on the concept of learning similarity but is a shortcoming for the task of reading comprehension. For the models to successfully reason about the questions they must have a more complex understanding of the underlying narrative. FairytaleQA (Xu et al., ACL 2022) furthers the discussion by creating a modified enriched version of NarrativeQA with educator based labeling for two question types, seven question themes, and who,what,where,why tagging which all are frameworks based in literacy education. This narrows the focus of the reading comprehension problem to grade school educational reading comprehension and applications. Progress is also made to demonstrate the success of BART fine-tuning but still shows opportunity for progress for challenging question types.

A MoE has been used to achieve higher performing results in other use cases by splitting the tasks then specializing models with the added benefit of reduced computational requirements (Cai et al., 2025). We pursued this simple yet powerful approach as a pseudo methodology due to our time limitations and relied on the question labeling provided by the dataset. We further noted this approach could yield better results than one model as shown in (Yang et al. 2024). Based on this work we also hypothesize that QWEN's long context capabilities may be advantageous for our dataset. We also pursued Flan-T5 small for our problem due to our resource limitations and its proven benchmark capabilities (Chung et al. 2024). Based on the paper we also see the success of the CoT method and believe the emphasis on reasoning could help our specific questions even with a zero shot approach required due to our dataset.

3 Methods

For our initial baseline and improvement over the baseline experiments, we use a Flan-T5 model for sequence to sequence text generation. We choose a Flan-T5 model because of its instruction based fine-tuning as well as its CoT fine-tuning mixture (Chung et al., 2022). We believe this is uniquely applicable to our problem as the instruction tuning will enable the request to answer a question based on context, and we anticipate the CoT fine-tuning will help with the tougher implicit questions that require some reasoning. For all sequence to sequence experiments, we set up the inputs to start with some instruction and context. We structure the input to the encoder as follows: *"Context: " followed by the story, then the instruction "Please answer this question: " followed by the question*. We use this structure because it aligns with how humans typically engage with stories–reading the narrative first, then answering questions based on it. While we will be reporting on ROUGE-L, BLEU, and BLEURT we also understand that each metric has its strengths and shortcomings. With this knowledge, we will define success in the following experiments as more than just improving the metric scores over the baseline, but also our own qualitative visual inspection and human evaluation of improvement in model performance.

3.1 Baseline

As a baseline we choose the Flan-T5 model with default parameters and no fine-tuning. This model was a good baseline model because it was able to perform our task with minimal data preprocessing to provide the question and context as input and produce responses without any tuning of the model. Initial inspection of model performance shows that there is a clear gap in performance on explicit and implicit questions. As shown in Table 1 you can see that, as expected, it is easier to answer explicit questions where answers can be found directly in the text versus implicit questions that require some reasoning or prediction.

3.2 Improvement over the Baseline

Our first attempt at improvement over the baseline was to fine-tune our Flan-T5 model over the entire dataset, including both explicit and implicit QA pairs. This does not satisfy the conditions of our pseudo MoE since this a singular model end to end, but is a good first look at the fine-tuned Flan-T5 performance across the different question types. For all model fine-tuning that follows, we use a 4-bit quantization of Flan-T5 together with a LoRA configuration for efficient parameter fine-tuning. For this experiment we fine-tune for 3 epochs with a small learning rate of 1e-5. This first fine-tuning experiment did not yield improved results as shown in Table 1. Except for BLEU, all evaluation metrics (ROUGE-L, and BLEURT) declined compared to the baseline. Analyzing the model performance, we do see that this is true for both explicit and implicit questions with the most noticeable improvement seen in the implicit answers; however, implicit answers still fall noticeably below explicit answers on all evaluation metrics. Furthermore, as we perform our qualitative analysis, we see an unexpected outcome of this experiment was that 32% of generated answers were simply “None of the above choices,” despite our dataset containing no multiple-choice questions. We hypothesize that this is an artifact of Flan-T5’s pre-training on multiple-choice formats, causing the model to default to this phrase when uncertain. We aim to mitigate this in future fine-tuning. The results of our first experiment further justifies the need for our proposed pseudo MoE architecture as the the requirements needed for answering explicit questions is fundamentally different than the requirements for answering implicit questions. Fine-tuning on both question types seems to be leading to a mediocre performance overall and individually.

Our next experiment was the first key attempt at our proposed architecture of the pseudo MoE where we fine-tune smaller specialized Flan-T5 models and have them answer their respective question types. Again, the intuition here is that these models can specialize in handling a specific question type and therefore their specialized performance together would yield better results than a singular model handling everything. For consistency we keep the same input structure to the encoder-decoder architecture with story, context, followed by the question but instead fine-tune two separate Flan-T5 models, one fine-tuned only on the explicit questions, and one fine-tuned only on the implicit questions. We fine-tuned for 5 epochs and a larger learning rate (1e-3) in an attempt to cut down on the “None of the above” output we saw before. Here we see clear improvement over the baseline across all three of our key metrics (ROUGE-L, BLEU, and BLEURT) for both explicit and implicit questions. Additionally, upon visual inspection we can see that we are no longer seeing “None of the above choices” in the generated answers. This is great to see that our proposed architecture is delivering improved performance as we expected; however, it’s clear that implicit question performance is still lagging behind explicit question performance. These two experiments were key in helping solve our problem because they enabled us to see where our proposed architecture was succeeding and where it needed improvement.

Since we see that our model is struggling more with implicit questions, the remainder of our experiments will be focused on improving the implicit question answering expert. As we have discussed, it makes sense intuitively that implicit questions are tougher than explicit questions because their answers cannot be found directly in the text. They will require some form of reasoning or prediction to come to the correct answer. To improve implicit question performance, we tested CoT prompting during inference. We opt to only implement CoT for inference, not fine-tuning, because we lack the data needed for fine-tuning with our dataset. However, given Flan-T5's CoT fine-tuning mixture we felt that this was a reasonable experiment to evaluate improvement on implicit question answering. In other words for implicit questions only, we are trying to force the model to reason about the question before giving a final answer. Unfortunately, the experiment was unsuccessful across all evaluation metrics since implicit scores dropped across the board. Rather than continue down this path, we will explore a different model architecture for our implicit expert.

To address the limitations of encoder-decoder models for implicit questions, we transitioned to a decoder-only architecture using QWEN-2-Instruct. We did attempt to work with Llama-Instruct models but were unable to move forward with them due to their size and computing restraints. We felt that a decoder-only model would improve implicit question answering because of how it generates text autoregressively. In other words, we felt that conditioning on previously generated tokens would give a better chance at reasoning and attending to the entire prompt. We felt that the decoder-only architecture would allow the required flexibility in responses needed to answer questions whose answers aren't found in the text. We decided to use the QWEN model architecture because of the scale of its high quality training on 7 trillion tokens, its long-context training that is crucial to understand the entire story context when answering questions, as well as the instruction based fine-tuning which includes instruction following and logical reasoning (Yang et al., 2024). Since this is a new architecture, we had to apply a different setup for fine-tuning and generation. In order to be more in line with how QWEN expects input, we apply a chat template and adjust the prefix instruction to specify answering the question with one sentence since our reference answers are all short one word to one sentence answers. After fine-tuning for 3 epochs and generating answers on the test set, we do see slightly lower metrics compared to our Flan-T5 implicit expert; however, our qualitative analysis shows that we see QWEN is producing longer, more conversational answers, but of higher quality than the Flan-T5 expert. We feel that even though the raw metrics degraded, QWEN on average is demonstrating a better comprehension of the story and its relation to the question compared to the Flan-T5 implicit expert.

After all of the above experimentation our final model architecture for the pseudo MoE is the fine-tuned Flan-T5 explicit model expert and the QWEN-Instruct implicit model expert. In the next section we will go into more detail about model experiment results and analysis.

4 Results & Discussion

FairytaleQA dataset is already split into train, validation, and test from its original creation (Xu et al., ACL 2022). We choose to retain the same split because the imbalance of question types is represented fairly as an approximate 75/25 split across all the groups. This also gave us the opportunity to compare our results against their best fine-tuned model test ROUGE-L score, 0.536. We speculate our best comparable score, 0.395, fell short because of resources and model size. Though they do not specify their BART model, BART-base has about 140M parameters whereas Flan-T5 small has 77M parameters.

The test data contains two references so we scored our metrics on both then used the max score in our average metric. Test questions offered the same reference for both fields 28% of the time, while for those that were different we observationally concluded that a fair amount of questions with different references indicate multiple correct but meaningfully different comprehensions of the story, such as in example 78. Therefore, we believed the max score was more appropriate than an average which may diminish the capabilities of our generated answers. We also recognize that because our max scores across metrics for one question could be from different references it may inflate our performance in some instances but we believe this to be minimal.

In the creation of our dataset there is sometimes a second human labeling for question type that conflicts with the primary assessment. We chose to use only the primary labelings for simplicity but see that in some instances implicitly labeled questions may be better described as explicit questions because the references are extracted from the context rather than abstractive. Unfortunately, as many as 80 out of our 253 implicitly labeled test data could be explicit questions and with a low n for the test population this could be degrading the evaluation of our question specific models architecture. This is an opportunity for improvement with a true MoE where a classifier for question type could decrease noise for the separately fine-tuned models.

We evaluated our models on the widely used ROUGE-L, BLEU, and BLEURT with each offering a unique perspective though we found qualitative evaluation to also be essential. We use ROUGE-L because it is the metric of choice in the original FairytaleQA and NarrativeQA work and it's a slightly flexible faithfulness metric which may be good for story based responses where answers might contain different additional phrasing in the answer. We use BLEU for its standardness and because of its stricter faithfulness since short responses are common, with 32% of test questions having primary references of 3 words or less. BLEURT is our most preferred metric especially for implicit questions because the abstractive responses can vary widely from our references while conveying a faithful answer. We also note, feeling based questions with one word feeling references account for 19 out of 20 of the most common one word references, which is at least 8% of the test data. These responses can be evaluated more granularly on semantic similarity using BLEURT as seen in example 60 where BLEURT does fairly well compared to the other metrics which give the response no credit.

Our best pseudo MoE model is our separately fine-tuned T5 model. On all metrics it outperforms our baseline with marginally better (0.021) results on ROUGE-L and (0.07) better on BLEURT and a substantial (0.248) improvement on BLEU (Table1). The explicitly fine-tuned T5 model also beats the baseline subsetted on explicit questions on all metrics. Additionally, the implicitly fine-tuned T5 model outperforms the implicit baseline averages, CoT, and QWEN on all metrics. With hands down the best numeric results, we also found that qualitatively our best model was able to correct a shortcoming of the baseline model by no longer using the response “None of the above”. Example 8 shows how the various models improved in that scenario. While it is progress to generate a response that is related to the context there is still a lack of success in our best model to produce a good result. Unfortunately it still favors extractive responses even though it has been tuned on only implicit questions. This tendency may be because the size of the training and validation implicit questions is small, only 2,447 total, or perhaps a larger T5 model with more trainable parameters would do a better job at learning our limited examples.

While we saw the most promising improvement over baseline on BLEU it appears to instill more confidence than perhaps it should. As seen in example 78, our best model seems to be doing well on BLEU when it doesn’t represent a good answer because the shorter response mimics context in the references but lacks the faithfulness of a correct comprehension of the story. Despite the success of the metrics for our best model we found the generated responses of QWEN were often preferable. QWEN responses tended to be longer than our best model in a more conversational style than the references. Example 78 also displays this scenario. Even though the reasoning in the QWEN response lacks some comprehension logic because the engagement in the story cannot be prevented, it demonstrates a more thoughtful interpretation of the story while scoring lower across all metrics.

The CoT implicit model was the poorest performing model across all metrics. Example 8 demonstrates the shortcomings in the CoT answers. The answers tended to be very lengthy and repetitive with numbered components and text mostly extracted from the context. While narratively the parts of the answer would be faithful to the references due to the extractive nature, overall the meaning, coherence and fluidity suffer to a degree congruent with the abysmal metrics. The other issue with CoT responses is unexpected brevity and lack of meaning like seen in Example 78 which behaves similarly to the baseline. This was a common mishap of the baseline as well as fine-tuned T5 experiments indicating the difficult nature of this problem and the essential role of fine-tuning for our data. This performance is not out of expectation for the CoT given the zero shot approach since we did not have labeled data to fine-tune with.

5 Conclusion

Our problem was to build a model that could mimic the reading comprehension skills required of grade school students with a focus on implicit and explicit questions. We were unable to see meaningful responses at a consistent level to provide useful functionality in real world applications like assisting educators with development and evaluation of classroom learning. We found that we were able to successfully answer many explicit questions accurately due to the extractive nature of those questions; however, the most challenging aspects of reading comprehension are mostly beyond the capabilities of the models we created. In future work a larger model may be more performant at learning these skills and we speculate that a true MoE could still be a beneficial approach.

6 References

[1] Ying Xu, Dakuo Wang, Mo Yu, Daniel Ritchie, Bingsheng Yao, Tongshuang Wu, Zheng Zhang, Toby Jia-Jun Li, Nora Bradford, Branda Sun, Tran Bao Hoang, Yisi Sang, Yufang Hou, Xiaojuan Ma, Diyi Yang, Nanyun Peng, Zhou Yu, and Mark Warschauer. 2022. Fantastic Questions and Where to Find Them: FairytaleQA – An Authentic Dataset for Narrative Comprehension. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 447–460, Dublin, Ireland. Association for Computational Linguistics.

[2] Tomáš Kočiský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. The NarrativeQA Reading Comprehension Challenge. Transactions of the Association for Computational Linguistics, 6:317–328.

[3] Yang, A., Yang, B., Hui, B., Zheng, B., Yu, B., Zhou, C., Li, C., Li, C., Liu, D., Huang, F., Dong, G., Wei, H., Lin, H., Tang, J., Wang, J., Yang, J., Tu, J., Zhang, J., Ma, J., Xu, J., Zhou, J., Bai, J., He, J., Lin, J., Dang, K., Lu, K., Chen, K., Yang, K., Li, M., Xue, M., Ni, N., Zhang, P., Wang, P., Peng, R., Men, R., Gao, R., Lin, R., Wang, S., Bai, S., Tan, S., Zhu, T., Li, T., Liu, T., Ge, W., Deng, X., Zhou, X., Ren, X., Zhang, X., Wei, X., Ren, X., Fan, Y., Yao, Y., Zhang, Y., Wan, Y., Chu, Y., Cui, Z., Zhang, Z., & Fan, Z. (2024). QWEN2 Technical Report. ArXiv, abs/2407.10671.

[4] Chung, Hyung Won, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li et al. "Scaling instruction-fine-tuned language models." *Journal of Machine Learning Research* 25, no. 70 (2024): 1-53.

[5] Cai, Weilin, Juyong Jiang, Fan Wang, Jing Tang, Sunghun Kim, and Jiayi Huang. "A survey on mixture of experts in large language models." *IEEE Transactions on Knowledge and Data Engineering* (2025).

7 Authors Contributions

Danielle ran the T5 models, experimented with fine-tuning and generation parameters (to achieve the final model), and experimented with BART (ultimately not included in this report) and wrote the deep dive analysis functions for model performance. She wrote the Background, Results & Discussions, and Conclusion sections of this report. In addition, she created the tables referenced in the appendix.

Dylan ran the fine-tuning T5 overall, small specialized fine-tuning (explicit vs implicit experts), chain-of-thought prompting, simple test time scaling (ultimately not included in this report but included as an appendix python notebook), Llama, and QWEN experiments. He wrote the Abstract, Introduction, and Methods sections of this report.

8 Appendix

Table 1

TEST DATA

Question Type	Model	ROUGE-L	BLEU	BLEURT
Implicit & Explicit n = 1007	Baseline (T5 out-of-the-box)	0.374	0.124	-0.583
	T5 fine-tuned together	0.308	0.255	-0.788
	T5 Explicit + T5 Implicit	**0.395**	**0.372**	**-0.513**
	T5 Explicit + QWEN Implicit	0.388	0.323	-0.534
Implicit	Baseline	0.183	0.020	-0.913
n= 253	QWEN	0.188	0.128	-0.944
	CoT	0.067	0.046	-1.397
	T5 Implicit	**0.218**	**0.323**	**-0.859**
Explicit	Baseline	0.439	0.158	-0.472
n= 754	T5 Explicit	**0.455**	**0.389**	**-0.397**

Example 60	
Story	the-false-prince-and-the-true
Question	how did the king feel after the gentleman did not want to marry the old woman ?
Reference 1	unhappy .
Baseline	frowned
RougeL/ BLEU/BLEURT	0.0 / 0.0 / -0.06

Example 8	
Story	happy-hunter-skillful-fisher
Question Type	Implicit
Context	between the sea realm and the earth there was no difference in the night of time , and the happy hunter found that three years went fleeting quickly by in this delightful land . the years pass swiftly when any one is truly happy . but though the wonders of that enchanted land seemed to be new every day , and though the sea king 's kindness seemed rather to increase than to grow less with time , the happy hunter grew more and more homesick as the days passed , and he could not repress a great anxiety to know what had happened to his home and his country and his brother while he had been away .
Question	what will the happy hunter do because he is homesick ?
Reference 1	try to go home .
Reference 2	runaway back home .
Baseline	None of the above choices .
T5	repress a great anxiety to know what had happened to his home and his country and brother
QWEN	he will go back to his homeland and family to find out what has happened to him.
CoT	1. the wonders of that enchanted land seemed to be new every day 2. the wonders of that enchanted land seemed to be new every day 3. he could not repress a great anxiety to know what had happened to his home and his country and his brother while he had been away 4. the happy hunter grew more and more homesick as the days passed , and he could not repress a great anxiety to know what had happened to his home and his country and his brother while he had been away

Example 78	
Story	the-dwarfie-stone
Question Type	implicit
Context	she agreed willingly . hiding her rosy face on his shoulder , she confessed that she had loved him from the very first day that she had seen him ; and ever since that moment she had determined that , if she could not we d him , she would we d no other man . for a little time they sat together , rejoicing in their new - found happiness . then earl paul sprang to his feet . " let us go and tell the good news to my mother and my brother , " he said . " harold may be disappointed at first , for i know , sweetheart , he would fain have had thee for his own . but his good heart will soon overcome all that , and he will rejoice with us also . " but the lady morna shook her head . she knew , better than her lover , what earl harold 's feeling would be ; and she would fain put off the evil hour . now , when the countess fraukirk had been away upon her wicked errand , strange things were happening at the castle at kirkwall . for harold , encouraged by his brother 's absence , offered his heart and hand once more to the lady morna . once more she refused him , and in order to make sure that the scene should not be repeated , she told him that she had plighted her troth to his brother . when he heard that this was so , rage and fury were like to devour him . mad with anger , he rushed from her presence , flung himself upon his horse , and rode away in the direction of the sea shore . while he was galloping wildly along , his eyes fell on the snow - clad hills of hoy rising up across the strip of sea that divided the one island from the other . and his thoughts flew at once to snorro the dwarf , who he had had occasion , as well as his step - aunt , to visit in bygone days . " i have it , " he cried . " stupid fool that i was not to think of it at once . i will go to snorro , and buy from him a love - potion , which will make my lady morna hate my precious brother and turn her thoughts kindly towards me . "
Question	what will happen after harold finds out about paul's and lady morna's engagement ?
Reference 1	harold will try to win lady morna 's love .
Reference 2	he will kill paul .
Baseline rougeL/BLEU/BLEURT	snorro 0.0/ 0.0/ -1.15
T5 rougeL/BLEU/BLEURT	the lady morna . 0.33/ 0.28/ -0.98
QWEN rougeL/BLEU/BLEURT	he will be furious and will try to break off the engagement before it can take place. 0.23 /0.13/-1.12
CoT	harold