Automated Extraction of Answer Candidates for Question Generation

Claudia Preda[1], Mihai Dascalu[1], Stefan Ruseti[1], Danielle S. McNamara[2]
[1] National University of Science and Technology Politehnica,
313 Splaiul Independentei, 060042, Bucharest, Romania
[2]Learning Engineering Institute, Arizona State University,
PO Box 871104, Tempe, AZ 85287
claudia.preda2307@stud.acs.upb.ro, {mihai.dascalu, stefan.ruseti}@upb.ro,
danielle.mcnamara@asu.edu

Abstract

Answering questions based on a reference text is a frequently employed comprehension assessment method that enables teachers to effectively and efficiently evaluate students. Various tools and methods were developed to tackle automated question generation, however, selecting valid answer candidates as a first step is less addressed. Thus, we introduce a solution built on top of FairytaleQA and tailored for training a DeBERTa-based model to classify the quality of each candidate to be part of a strong answer-question pair. First, we extract answer candidates by syntactically parsing the context (i.e., selecting text spans from the reference text based on the nodes in the constituency tree); then, questions are generated for the extracted candidates using a pre-trained LLM model on this task. Next, we assess a candidate's quality by relying on another fine-tuned model's capability to answer the previously generated question for that candidate. This enables us to categorize answers using a four-class system: very good, good, average, and unusable. A significant advantage of our method is that the encoder classifier can score all potential answer candidates in a single inference step for the entire context. We compare our selection against both the answers from explicit questions in the original dataset and a fine-tuned LLM for answer selection using an Elo ranking system. In addition, we propose three strategies based on semantic similarity and text position to ensure coverage and diversity of candidates' selection.

Keywords: answer extraction, large language models, constituency tree, single inference step

1. Introduction

Reading comprehension is an essential skill that develops critical thinking beyond the understanding of written texts. The concept is thoroughly discussed in the educational literature (Oakhill et al., 2014), stressing the importance of developing efficient methods to help students improve their skills in comprehending what they read. Question-answering is a frequently employed educational approach that helps students develop strategies for understanding texts by formulating and answering questions through practice. Its primary goal is to foster the development of comprehension strategies through the process of generating and responding to questions. This is not a skill acquired instinctively, but it is crucial for understanding the text's structure and synthesizing written information.

Educational professionals face time constraints when evaluating their students due to the complexity of creating question-answer pairs that fit diverse contexts and scenarios. One popular, but time-consuming, example is a strategy called Reciprocal Questioning, which involves two readers taking turns asking each other questions (Manzo, 1969). Relying on artificial intelligence technology has proved an efficient method for improving reading comprehension (Hidayat, 2024) and for personalized development (Ademola, 2024). This creates the ideal environment for pursuing additional methods to make the process more efficient and achieve better results.

Answer selection or generation from a support text has emerged as an enhancement for two other important tasks in Natural Language Processing: question answering and question generation. Most often, the methods used to extract the candidates are named entity recognition, keyword extraction, summarization, or paraphrasing. In most cases, these methods provide a limited set of answers extracted from the given text.

Despite its recent popularity, generating question-answer pairs remains challenging because it is difficult to emulate human-like reasoning and ensure that the most suitable options are considered. As such, the diversity of selected answers from the reference text and their adequacy for creating relevant questions remain important aspects to be tackled.

Hence, we propose a novel approach to address these challenges and to facilitate the extraction of text-based answer candidates in a single inference step for each context fed into the model. Our method shifts the focus to answer selection from a pool of candidates generated by syntactic parsing of the context, enabling us to study every structure available without external intervention. The next step is to classify the selected text spans based on their capacity to form answer-question pairs. The final candidates are ranked according to the probability of being labeled as part of the very good class.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 11141–11151
11–16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

This approach has two goals: offering teachers flexibility to quickly generate a suite of question-answer pairs for a given text and creating synthetic data for future use in improving related tasks, such as feedback-aware approaches or exploring the difficulty of a pair.

To argue for our method's effectiveness, we compared it with the human-selected answers from the original dataset and two other methods of generating candidates: a fine-tuned LLM and named entities. Further, we evaluate the question-answer pairs using an Elo ranking system (Elo, 1967), where the answers are obtained via two methods: our proposed solution and a fine-tuned model. We compare the models against each other and against human-generated pairs from the used corpus.

As main advantages of this approach, we identify the diversity of the candidates and, implicitly, of the generated questions, as well as the control over the selection process. For each paragraph, we can obtain all candidates at once and ensure that the desired answer is within the provided context. As such, we provide three selection strategies based on text similarity and the positions of potential answer candidates within the provided context.

Our main contributions are as follows:

- Introduce an encoder classifier that scores all potential answer candidates in a single inference step for the entire context.
- Provide a diverse pool of explicit candidates that can form question-answer pairs, without external feedback.
- Introduce three pre-defined strategies (i.e., top k filtered, k-means++, and top k sentences) that ensure coverage of different options for selecting candidates in a given text.

We release our code as open source on GitHub.

2. Related Work

2.1. Question Generation and Answer Selection

Automated question generation is a popular task in the educational field, playing an important role in reading comprehension assessment. The methods used to automatically generate open-ended questions evolved from RNNs (Bi et al., 2020) to Transformers (Xu et al., 2022), and now leverage the capabilities of Large Language Models (Li and Zhang, 2024).

Strategies varied from using patterns and templates (Ali et al., 2010) to summarization, named entity recognition, part-of-speech tagging, or based on extracted concepts, as the relationship between text elements. For example, Zhao et al. (2022) used the type of question distribution in a paragraph, such as causal relationships, outcome resolution, and prediction, to control the summarization of events in the given text. The summary is passed to a transformer-based model to generate a question. Another approach by Li and Zhang (2024) leveraged LLMs to generate a plan based on annotated labels for questions in a dataset containing contextually important events. The plan helps another model in building a suitable question for the source text.

Willis et al. (2019) focused on answer extraction from the reference text and on evaluating the obtained pairs. Their technique used an encoder-decoder model to generate key phrases based on part-of-speech tagging and named entity recognition. The conducted experiments found that this method was better suited to factual questions, producing key phrases composed of a single word.

On a similar note, Yao et al. (2021) generated multiple answer-question pairs from a text using heuristic-based rules for answer selection. The methods used to extract the candidates include noun chunks, named entities, and event descriptions. The author ranked the output by evaluating each question-answer pair against the samples from the original dataset. Their results showed that the probability of the system performing better increases with a larger pool of potential top candidates.

Other approaches used similar techniques to create synthetic questions for the dataset augmentation (Lee et al., 2023; Nagumothu et al., 2023). Both studies processed the reference text to express the meaning more concisely and improve the quality of the selected answer candidates. Lee et al. (2023) followed a more traditional approach, summarizing the text and extracting answers via named-entity recognition. They concluded that this method produced a more specific question because a common topic or fact typically relates to entities in the summary. Nagumothu et al. (2023) took a different approach by paraphrasing the initial context and using Open Information Extraction to select triplets that represent facts. They argued that their method ensured well-formed and effective training data. Both techniques improved the question-answering task on the used baselines.

Recently, the strategy has shifted towards prompt engineering applied to LLMs. For example, Wang et al. (2025) highlighted the importance of providing examples similar to those expected, resulting in better performance in education-related question answering. However, finding the right prompt for the diversity of texts a teacher can work with is not straightforward.

2.2. Using LLMs as a Judge

One of the most time-consuming aspects of evaluating a model's performance is the shortage of human evaluators. This type of evaluation is complex and requires familiarity with the research topic. Until recently, traditional metrics such as BLEURT (Sellam et al., 2020) and ROUGE (Lin, 2004) were used to assess performance. However, their drawbacks include their inability to capture subtler nuances and their reliance on similarity and lexical overlap.

As a result, the idea of using larger LLMs to evaluate the output of other automated solutions has become more popular. Zheng et al. (2023) proposed a framework for performance assessment using the latest models at that time, such as GPT-4. They showed that LLM-as-a-judge achieves 80% agreement with human evaluation, comparable to agreement between humans alone.

The available approaches mainly rely on prompting engineering for evaluation because they do not require additional training. Huang et al. (2024) analyze the benefits of fine-tuning models for this task, but conclude that while the performance increases for domain-specific tests, overall, general LLMs have more benefits. For open-ended questions related tasks, the evaluation method is pairwise comparison, meaning the judge should choose the better option from the outputs obtained from different sources.

One major drawback of using an LLM as a judge is that its reasoning and outputs can be biased. Ye et al. (2024) performed an analysis of different types of biases, such as position, verbosity, and distraction. Their paper also emphasized the importance of carefully crafting the prompt and encouraging the model to provide an explanation and its reasoning as mitigation.

3. Method

Our main task is to select multiple text sequences in a single step for a given context and rank them by their potential to generate question-answer pairs. Our workflow is presented in Fig. 1, with the main steps (i.e., data generation, classification, and inference to select the best pairs based on the chosen strategy) described in the following sub-sections.

3.1. Data Generation

For this study, we used the FairytaleQA dataset (Xu et al., 2022), designed for narrative comprehension by experts in the education sector. FairytaleQA comprises 10,580 explicit and implicit questions split into train, validation, and test partitions. Since our method is scoped to the reference text and has no external knowledge, we kept only the explicit questions from the dataset. Also, we maintained the original dataset split for all steps described in the following sections.

The first step towards building the input expected by the classification model is to create the constituency tree for each story section in the original dataset - here, we used the Berkeley Neural Parser (Kitaev and Klein, 2018). As we explored all nodes from the tree (i.e., leaves as words and all intermediary nodes), we obtained a large corpus containing over 1M entries for the training split and around 100k entries for each validation and test set. Although some nodes are obviously bad answers, we decided to keep all of them in the dataset and not add further filtering, since the model should learn to distinguish between suitable and unusable sequences.

The following two steps require fine-tuning a Large Language Model (in our case, Llama 3.2 3B (Meta, 2024), chosen for its performance and efficiency) on the original FairytaleQA dataset for question generation (QGEN) and question answering (QA). The first model generates a question for the context and the extracted candidate. The QA model estimates answer quality by computing the loss (sum of negative log probabilities) of the answer for each question-answer pair, with the question and the reference text being the input, while the answer candidate is the target.

Next, the QA loss is used as a base to label our candidates into four categories, as follows: very good, good, average, and unusable. The labels are assigned by sampling the training dataset and correlating the computed score with what is traditionally considered a good and a bad candidate. Answers we know generate well-formed questions are usually one of the following functions in a sentence: subject, direct object, or attribute. They are correlated with who-what-how types of questions and are easy to identify. In contrast, stopwords are unusable candidates. Next, we ordered all candidates for a given context by their loss and found a pattern: illogical pairs were more often present. Hence, the loss intervals empirically set for each label on the FairyTaleQA dataset are: *very good* - $[0, 10]$, *good* - $(10, 13]$, *average* - $(13, 15]$, and *unusable* - $(15, \infty)$.

3.2. Building the Classification Model

As our main purpose is to generate multiple answers for the same text, we decided to use a classification model that learns to differentiate between usable answers (labeled as very good, good, and average) and unusable text selections. The reason for this decision is that computing losses for all parts obtained by the syntactic parsing of the reference text is very time-consuming due to the

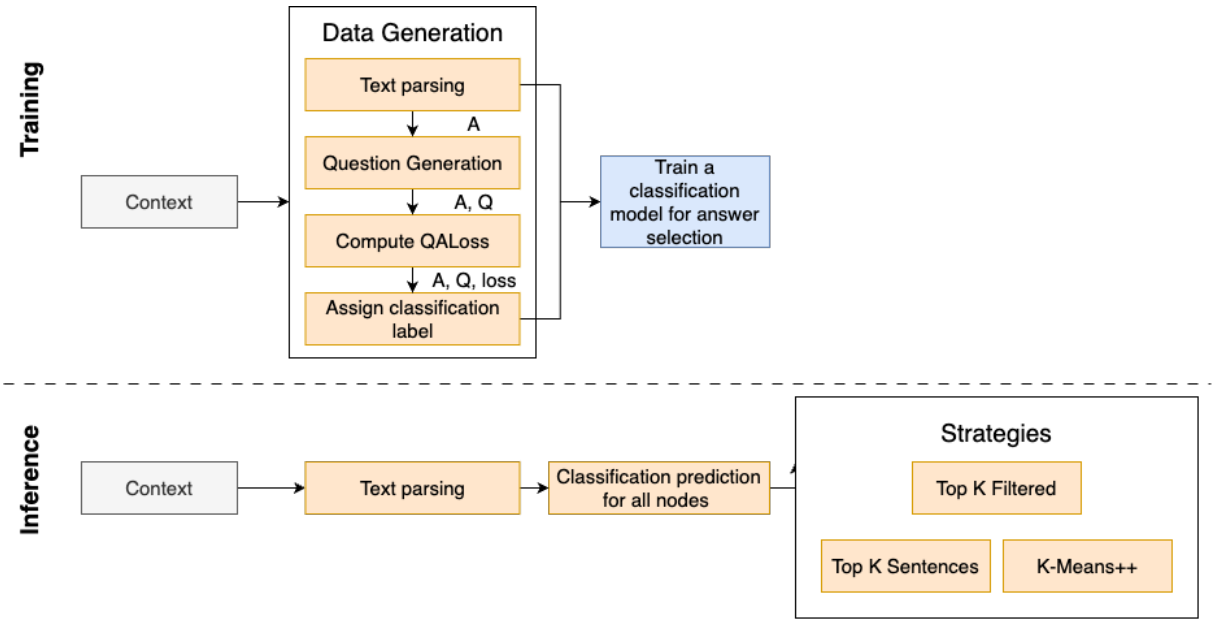

Figure 1: Solution workflow schema representing the difference in steps between the training phase and the single inference step.

question-generation process. Thus, we aim to build a quicker model based on the obtained data.

The classification model receives as input the context and a mask indicating the positions where each extracted sequence occurs in the context. To capture the relationship between the reference text and the candidate, we use and encoder-only transformer model to process the context, and compute the average hidden representation of the sequence using the mask. The classification output is computed with a feed-forward network with 3 layers. This architecture, shown in Fig. 2, enables efficient training and inference by passing the text through the encoder only once and computing the classes for all candidate answers in a single step. For this scenario we build two versions of the classification model based on DeBERTaV3 (He et al., 2020, 2021) and ModernBERT (Warner et al., 2024), which are the standard for encoders. DeBERTaV3 is more accurate for traditional tasks, such as classification, and uses the data more efficiently during training, while ModernBERT is a fast solution that can handle large contexts (Antoun et al., 2025).

For the training setup, both classifiers were trained on the same data for seven epochs using Adam optimizer and starting from the HuggingFace[1] base encoders. While ModernBERT accepts a larger input up to 8k tokens, the FairytaleQA dataset has shorter texts that fall under the 512 tokens threshold. We chose the different learning rates based on the architecture type: 1e-4(DeBERTaV3) and 2e-5(ModernBERT). To note that varying the learning rate for ModernBERT has not improved the performance, nor increasing the number of epochs.

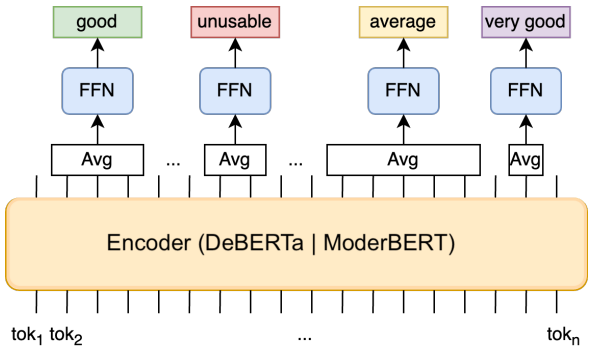

Figure 2: The classification model architecture that receives as an input the context and a list of possible answers to be classified in one of the classes: 1 - very good, 2 - good, 3 - average, 4 - unusable.

3.3. Inference of Best Answers

In the inference step, the context is parsed through the same method described in 3.1 section, resulting in all possible sequences from the constituency tree being classified, all at the same time, as *very good*, *good*, *average*, or *unusable*. To further select the best candidates for the reference text, the sequences are sorted by their probability of being a very good answer (the softmax layer output for this label). From these, we can select answers based on a desired strategy.

[1] https://huggingface.co/

Top 5 Filtered (not similar)	Top 5 Sentences (one candidate per sentence)	K-Means++ (text distance, semantic distance)
So the old woman continued, ' the most beautiful woman in the whole world is the daughter of the queen of the flowers, who has been captured by a dragon. If you wish to marry her, you must first set her free, and this i will help you to do. I will give you this little bell : if you ring it once, the king of the eagles will appear; if you ring it twice, the king of the foxes will come to you; and if you ring it three times, you will see the king of the fishes by your side. These will help you if you are in any difficulty. Now farewell, and heaven prosper your undertaking. ' she handed him the little bell, and there disappeared hut and all, as though the earth had swallowed her up.	So the old woman continued, ' the most beautiful woman in the whole world is the daughter of the queen of the flowers, who has been captured by a dragon. If you wish to marry her, you must first set her free, and this i will help you to do. I will give you this little bell : if you ring it once, the king of the eagles will appear; if you ring it twice, the king of the foxes will come to you; and if you ring it three times, you will see the king of the fishes by your side. These will help you if you are in any difficulty. Now farewell, and heaven prosper your undertaking. ' she handed him the little bell, and there disappeared hut and all, as though the earth had swallowed her up.	So the old woman continued, ' the most beautiful woman in the whole world is the daughter of the queen of the flowers, who has been captured by a dragon. If you wish to marry her, you must first set her free, and this i will help you to do. I will give you this little bell : if you ring it once, the king of the eagles will appear; if you ring it twice, the king of the foxes will come to you; and if you ring it three times, you will see the king of the fishes by your side. These will help you if you are in any difficulty. Now farewell, and heaven prosper your undertaking. ' she handed him the little bell, and there disappeared hut and all, as though the earth had swallowed her up.

Figure 3: Example of how selection strategies can be applied to assure diversity and text coverage.

To emphasize the solution's practical usability, we defined a set of strategies, each focused on a different aspect. All strategies can be applied in one single step for a given text and follow the approach: the reference text is syntactically parsed; the obtained sequences are processed by the DeBERTa Classification model and assigned a label; the selected candidates are ordered by the probability score of being in class 1 (very good answers); selecting a strategy for selection and applying it. As such, we introduce three strategies: Top K Filtered, K-Means++, and Top K Sentences, all depicted in Figure

For *Top K Filtered*, the first sequence after sorting the candidates is chosen as a reference. Then, we compute the BLEURT score to measure the similarity between the reference sequence and the other candidates. We then select the top *K* responses: the reference sequence and the top K = 1 less similar answers.

K-Means++ applies an approach similar to the K-Means++ initialization of centroids while using as a criterion a score that represents the average between the semantic distance and the text distance. For the semantic distance, we use BLEURT to compute the similarity and select the least similar sequences. For the text distance, we use the number of characters between two candidates, divided by the length of the context. We select the furthest candidates that are less similar to the previously selected answer candidates.

The last strategy, *Top K Sentences*, selects the answer with the highest probability of being a very good candidate among the most important *K* sentences in the context. We use the Sentence Transformers module (Reimers and Gurevych, 2019) and the Mpnet model (Song et al., 2020) to compute the similarity matrix between the sentence embeddings. We use PageRank as an extractive summarization method to compute the final importance score (Page et al., 1999) per sentence.

4. Results

We analyze the performance of the classification model in the following scenarios:

- We use the automatically computed classes with the QA loss as ground truth. We target the classification's overall performance based on the reference data.
- We compare our generated answers with the actual answers selected by experts for the explicit linguistic questions from the original dataset. Also, for reference, we add a fine-tuned Llama 3.2 3B for answer selection. We apply traditional metrics for this evaluation step (i.e., BLEURT (Sellam et al., 2020) and ROUGE (Lin, 2004)).
- We use Qwen3-Next-80B-A3B-Thinking (Yang et al., 2025; Team, 2025) as a judge for question-answer pairs and determine the better pair for the context. We chose a model from a different family than LLama to avoid a bias towards our Llama Answer Selector model which was fine-tuned on Llama 3.2 3B. On top of that, this is a model tailored for reasoning and the Mixture-of-Experts architecture allows it to be fast, while still being one of the benchmarks large language models. Performance was evaluated using an Elo ranking system (Elo, 1967). This system is used to calculate the score after two-players games relative to their current ranking and difference

in skills. After each game, the ratings are updated based on the difference in scores, resulting in more points for a win against a better-ranked opponent and fewer points for a loss to a significantly weaker opponent.

All experiments are performed on the FairytaleQA dataset.

4.1. Ground Truth Comparison

The DeBERTa Classification model achieves an overall micro-F1 Score of 0.727 while taking into account all possible answers. For a more realistic usage scenario, we also evaluated the *top-10* best-scoring unique candidates. Out of the top-10 selected candidates, 74% were considered *very good*, 20.6% *good*, 2.7% *average*, and 2.8% were *unusable*. For ModernBERT Classification model, we obtained a micro-F1 Score of 0.697. ModernBERT-based model selection for the top-10 candidates has the following composition: 19.29% *very good*, 17.41% *good*, 28.91% *average*, and 34.39% *unsusable*.

A Kruskal-Wallis test for the DeBERTa Classification model indicated that there was a significant difference in *loss value* across *4 [category],* $X^2(3, N = 122876) = 193203.536, p < .001$. The mean rank of the loss function was *15893.69* for category 1 (very good), *30976.75* for category 2 (good), *42342.26* for category 3 (average), and *74227.36* for category 4 (unusable). In comparison, for the ModernBERT Classification model, the Kruskal-Wallis test reported *loss value* across *4 [category],* $X^2(3, N = 122876) = 1316.0456, p < .001$. The mean rank of the loss function was *28941.19* for category 1 (very good), *30585.85* for category 2 (good), *53115* for category 3 (average), and *61863.38* for category 4 (unusable). This shows us that the DeBERTa based model is the better classifier because it clearly distinguishes between each classes, while the ModernBERT has a higher overlap between classes. In Fig. 4 and Fig. 5 and we can observe the distribution of the loss values for each category for both solutions, which supports the conducted analysis. On top of that we can observe that ModernBERT struggles with the 'average' class, with over 60% of this class ground truth entries being classified as 'unusable' by the model. Based on the above we conclude that DeBERTa Classification model is more suitable to our task and we will analyze it further, however we will include the comparisons for ModernBERT classifier for visualizing its performance.

4.2. Metrics Comparison

In the second experiment, we evaluated the model's performance on the top-10 candidates against human annotations in the original dataset and against other methods for selecting spans from the given text. The first baseline is represented by the list of all named entities from the context, extracted with `spaCy's NER model`[2]. The second baseline is a fine-tuned Llama model trained on the FairytaleQA dataset to generate all answers simultaneously; as such, no threshold on the number of retrieved answers was imposed.

We apply the same evaluation to each model, generating tuples of the form (label, candidate, score) for all combinations of ground-truth answers and candidates at the context level. The score is either BLEURT or ROUGE (longest common sequence). Then, we constructed a weighted graph to compute the maximum-weight matching, ensuring that each entry is used exactly once. We compute the average score for each context across the dataset. As seen in Table 1, our classification model obtained the best results, still comparable to the fine-tuned Llama 3.2 3B model.

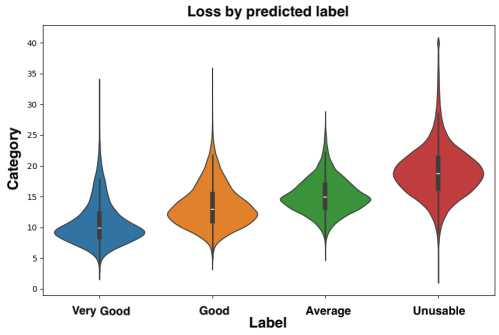

Figure 4: Distribution of loss by DeBERTa Classification model predicted label with: 1(Very Good), 2(Good), 3(Average), 4(Unusable).

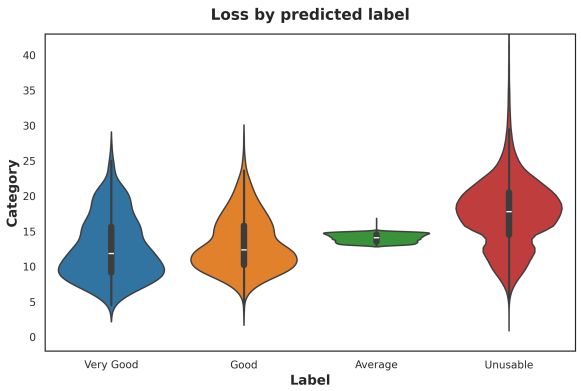

Figure 5: Distribution of loss by ModernBERT Classification model predicted with: 1(Very Good), 2(Good), 3(Average), 4(Unusable).

In Fig. 6, an example from the Fairytale dataset illustrates the capabilities of the proposed solutions.

[2] https://spacy.io/models/en

METHOD	BLEURT	ROUGE-L
NER Extraction	0.096	0.083
Llama Answer Selector	0.657	0.326
DeBERTa Classification	**0.663**	**0.878**
ModernBERT Classification	0.408	0.467

Table 1: BLEURT and ROUGE-L Scores (bold marks the best result).

RANK	MODEL	SCORE
1	Ground Truth Dataset	1564.50
2	DeBERTa Classification	**1512.06**
3	Llama Answer Selector	1455.73
4	ModernBERT Classification	1425.33

Table 2: Overall Elo Rankings (bold marks the best results besides the human ground truth).

It can be observed that most answers fall in the *very good* category, and the pairs are comparable with the example from the original dataset. Also, we notice that DeBERTa Classification selects answers that produce more varied types of questions compared to Llama Answer Selector, which extracted mostly answers for "What" type questions. To note that we ran the fine-tuned model ten times and obtained only *four* different answers, from which one was not present in the support text, and it was a wrong interpretation of the context. The selected answer *"he was afraid to reject her"* implies that the prisoner was afraid of the woman who offered to help him, instead of the prisoner being afraid of death (from the context: *" the full horror of his coming death rushed upon the young man"*). Thus, despite our model being limited to selecting only answers specifically mentioned in the context, we ensure it is more reliable and avoids contributing to irrelevant pairs.

4.3. Elo Ratings Comparison

For this experiment, we first used the Qwen3-Next-80B-A3B-Thinking model to perform a pairwise comparison between question-answer pairs, where the answer is selected by our DeBERTa Classification model, by the fine-tuned Llama Answer Selector model, or from the original Fairytale dataset. The questions are generated using a finetuned Llama 3.2 3B, except for those already present in the dataset. We compare the two models against each other and both against the human-generated pairs in the dataset. For the comparison between DeBERTa Classification and Llama Answer Selector, we select the top 10 answers and their questions for each context. For each such pair, we select a new answer using Llama Answer Selector and generate a question. For the comparison between our model and the ground truth, we form all combinations of each top-10 pair per context with the pairs in the dataset for the same context. Lastly, for each pair in the dataset, we select an answer using the fine-tuned model and generate its corresponding question. We also replicate the experiments for the ModernBERT Classification model with the addition that we also compare it against the DeBERTa Classification model.

To mitigate biases and limitations that LLM displays when asked to act as a judge, we follow the recommendations from Ye et al. (2024). Thus, we randomly permute the pairs given as input to the model, thereby avoiding the tendency to select the first answer, which would otherwise display a clear preference for one option. The model is instructed to provide its reasoning for the choice and to remain impartial.

The judge's task is to establish which question-answer pair is better suited for the given text, considering the following criteria: *the answer and the question are based on the given text; the answer addresses the question; the pair assesses understanding rather than simple recall*. The model is allowed to declare a tie only when the decision is difficult; otherwise, it should declare a win if the first pair is better or a loss if the second is better.

To evaluate the results, an Elo ranking system (Elo, 1967) was built to assess LLM performance. The algorithm has been proven to be efficient and it was widely studied to be optimized for LLM evaluation, for example Boubdir et al. (2024) compared the Elo ratings to human evaluation and drew conclusions about efficiency, especially about adjusting parameters to reduce rating fluctuations and the need to adjust the K factor (maximum possible adjustment per game) for rapid convergence for clear winners. Similarly, a system based on Elo (Chiang et al., 2024) is used for the Chatbot Arena LLM Leaderboard[3].

For our experiment, we used an initial rating of 1500 and a K-factor equal to 32. We compared the models and also simulated a competition of random matches among all three models.

In Table 2, we observe that our model ranks second overall, with pairs from the original dataset perceived as better. This is expected because our top 10 selected answers may also include easier options that do not always target the text understanding. Our model achieves a better score than the fine-tuned Llama, which also has as a drawback the fact that it can extract answers that are not in the given context. The ModernBERT Classification does not perform well compared to the other solutions and to DeBERTa Classification, supporting the other experiments conducted in this paper.

[3] https://huggingface.co/spaces/lmarena-ai/chatbot-arena-leaderboard

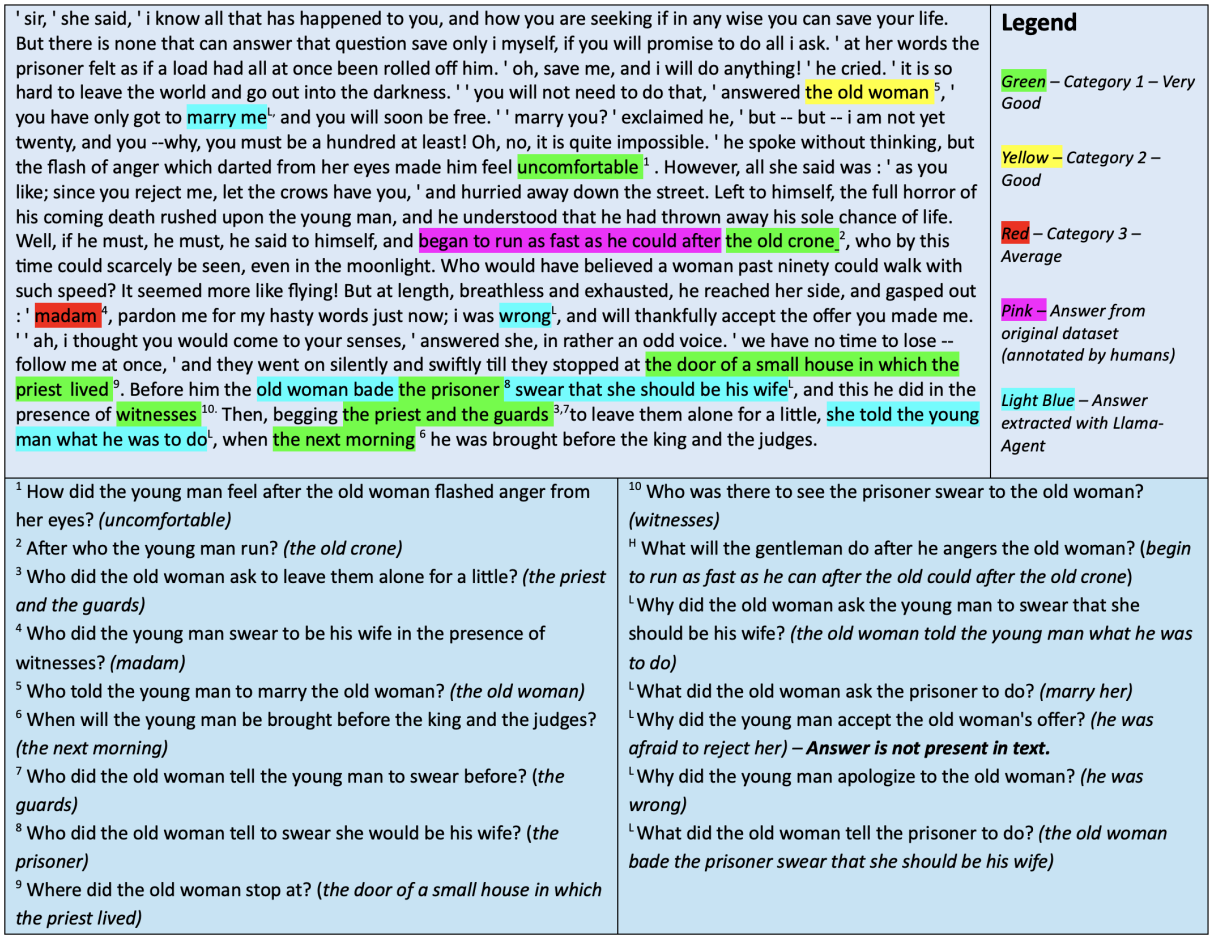

Figure 6: DeBERTa Classification model results compared with the original question-answer pair annotated by humans in the Fairytale Dataset for the given context and with the Llama fine-tuned model. Note: Questions 1-10 are generated for the answers selected by the encoder model (DeBERTa Classification), H is used for the original question, and L for the answers extracted using the fine-tuned Llama Answer Selector.

5. Discussion

The classification model and the empirically chosen label ranges have proven efficient at generating question-answer pairs using only the information in the context. We chose to split the answers into multiple categories to maintain a ranking of usable answers and to illustrate that the current solution will predominantly select answers from the best class. However, this study focuses more on the possibility of selecting diverse answers, rather than strongly differentiating between the three usable classes. On the performance gap between the two classification models, we can identify that ModernBERT is affected by the labels imbalance in the dataset, as the 'unusable' class has more entries than the others, representing around 55% from the training set. It is a less sensitive model to changes being suitable for larger contexts and being optimized to be fast. Thus, for this classification task we need a more precise model and DeBERTa's Disentangled Attention (separating content from relative position) is a major advantage.

A couple of factors impact our overall performance. First, the selected spans might contain duplicate answers as the current method is not able to distinguish between subtle changes in meaning, such as *("under the hill," "the hill")* or *("in a while," "a while")*, assigning them the same label and making abstraction of the sense. This happens because we use average hidden states to represent a sequence, which does not change significantly when adding or removing a token. One solution is to use a more complex aggregation function on top of the hidden states, like a CNN or LSTM. In Fig. 6, we illustrate this scenario for answer 7, which is part of answer 3, leading to similar questions. This is also the reason for selecting a classification model rather than a regression model, since the latter was unable to learn subtle differences. When adding more strategies for answers selection, this issue is solved, but improving this aspect results in being able to select a greater diversity of candidates with

less effort.

Second, when we established the correspondence between the loss score and the assigned label, we identified an anomaly: some of the longer spans that could form a strong pair have loss scores in the unusable category. One example is the following pair *("spreading kingdom," "his beautiful palace and all its wonders," and "his power which none disputed throughout the whole sea," "What did the dragon king have?")*, which was validated to have sense in the given context. However, there is no clear pattern, and including these outliers would compromise the model's overall performance, as most unusable spans have a similar score.

6. Conclusions and Future Work

Our method successfully and efficiently extracts candidates from a reference text comparable to the human-annotated data in the original dataset. Despite similar performance to the Llama fine-tuned model, we argue that the proposed model is lighter (an encoder) and faster, processing all spans within a single inference step. Furthermore, the fine-tuned model lacks diversity, repeatedly selecting the same answer from the given text. One aspect that highlights the model's strengths is its capability to define diverse strategies for controlling the selection of possible answers.

For future work, we aim to improve the metrics that best define the relationship between a given text and what constitutes a good question-answer, without relying on human annotations. Also, we would like to explore how to generate well-balanced and diverse question-answer pairs. On the encoder choice, we would like to explore further the capabilities of ModernBERT and further fine-tuned the model, due to its advantage of handling large contexts and faster inference, which makes it suitable in low resources environments.

7. Limitations

We identify the lack of human evaluation on the final test data as the main limitation. While an important method of evaluation, it is not always available and can be time-consuming. For generating question-answer pairs, human evaluation remains the most reliable option; however, comparing results with LLM performance and state-of-the-art models is a good compromise that speeds up the process, especially when the solution is intended to be further integrated to boost the performance of other tasks. While this study utilizes the Fairytale dataset to develop the solution and evaluate its performance, our approach can be easily applied to other datasets because it is independent of the data. Ideally, we only need to fine-tune the large language models for each task (question generation, question answering, and answer selection); however, a base model would be enough for quickly integrating the proposed solutions with other datasets.

While developing a methodology to evaluate the classification model, we faced the challenge of finding a suitable metric to capture the complex relationships between outputs and targets. If the selected candidate is not among the references, it can still form a good question-answer pair. Hence, one limitation is the lack of automatic evaluation, a topic also addressed by Yao et al. (2021).

8. Ethical Considerations

This study is built exclusively on top of the FairytaleQA dataset (Xu et al., 2022), a publicly available educational corpus. The dataset has been designed for the text comprehension task and it is sourced from the Project Gutenberg website which abides by the copyright law[4]. The data has been collected and released by the authors with the purpose of research reuse. No additional human data were collected. The dataset does not contain any personally identifiable information.

All experiments were conducted using open-source models and tools under their respective licenses. The study does not contain any sensitive or private user data and no human subjects were involved. Model outputs were analyzed solely for research evaluation purposes.

9. Acknowledgment

This research was supported by the project "Romanian Hub for Artificial Intelligence - HRIA", Smart Growth, Digitization and Financial Instruments Program, 2021-2027, MySMIS no. 351416.

10. Bibliographical References

OE Ademola. 2024. Reading strategies in the ai age: Enhancing comprehension and engagement with advanced technologies. In *Proceedings of the 38th iSTEAMS Multidisciplinary Bespoke Conference*.

Husam Ali, Yllias Chali, and Sadid A Hasan. 2010. Automatic question generation from sentences. In *Actes de la 17e conférence sur le Traitement Automatique des Langues Naturelles. Articles courts*, pages 213–218.

[4] https://www.gutenberg.org/policy/terms_of_use.html

Wissam Antoun, Benoît Sagot, and Djamé Seddah. 2025. Modernbert or debertav3? examining architecture and data influence on transformer encoder models performance. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 3061–3074.

Sheng Bi, Xiya Cheng, Yuan-Fang Li, Yongzhen Wang, and Guilin Qi. 2020. Knowledge-enriched, type-constrained and grammar-guided question generation over knowledge bases. *arXiv preprint arXiv:2010.03157*.

Meriem Boubdir, Edward Kim, Beyza Ermis, Sara Hooker, and Marzieh Fadaee. 2024. Elo uncovered: Robustness and best practices in language model evaluation. *Advances in Neural Information Processing Systems*, 37:106135–106161.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot arena: An open platform for evaluating llms by human preference.

Arpad E Elo. 1967. The proposed uscf rating system, its development, theory, and applications. *Chess life*, 22(8):242–247.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Muhamad Taufik Hidayat. 2024. Effectiveness of ai-based personalised reading platforms in enhancing reading comprehension. *Journal of Learning for Development*, 11(1):115–125.

Hui Huang, Xingyuan Bu, Hongli Zhou, Yingqi Qu, Jing Liu, Muyun Yang, Bing Xu, and Tiejun Zhao. 2024. An empirical study of llm-as-a-judge for llm evaluation: Fine-tuned judge model is not a general substitute for gpt-4. *arXiv preprint arXiv:2403.02839*.

Nikita Kitaev and Dan Klein. 2018. Constituency parsing with a self-attentive encoder. *arXiv preprint arXiv:1805.01052*.

Seongyun Lee, Hyunjae Kim, and Jaewoo Kang. 2023. Liquid: A framework for list question answering dataset generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 13014–13024.

Kunze Li and Yu Zhang. 2024. Planning first, question second: An llm-guided method for controllable question generation. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 4715–4729.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Anthony Vito Manzo. 1969. *Improving reading comprehension through reciprocal questioning*. Syracuse University.

AI Meta. 2024. Introducing meta llama 3: The most capable openly available llm to date, 2024. *URL https://ai. meta. com/blog/meta-llama-3/. Accessed on April*, 26.

Dinesh Nagumothu, Bahadorreza Ofoghi, Guangyan Huang, and Peter W Eklund. 2023. Pie-qg: Paraphrased information extraction for unsupervised question generation from small corpora. *arXiv preprint arXiv:2301.01064*.

Jane Oakhill, Kate Cain, and Carsten Elbro. 2014. *Understanding and teaching reading comprehension: A handbook*. Routledge.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford infolab.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur P Parikh. 2020. Bleurt: Learning robust metrics for text generation. In *Proceedings of ACL*.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. *Advances in neural information processing systems*, 33:16857–16867.

Qwen Team. 2025. Qwen3 technical report.

Lili Wang, Ruiyuan Song, Weitong Guo, and Hongwu Yang. 2025. Exploring prompt pattern for generative artificial intelligence in automatic question generation. *Interactive Learning Environments*, 33(3):2559–2584.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas,

Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference.

Angelica Willis, Glenn Davis, Sherry Ruan, Lakshmi Manoharan, James Landay, and Emma Brunskill. 2019. Key phrase extraction for generating educational question-answer pairs. In *Proceedings of the sixth (2019) ACM conference on learning@ scale*, pages 1–10.

Ying Xu, Dakuo Wang, Mo Yu, Daniel Ritchie, Bingsheng Yao, Tongshuang Wu, Zheng Zhang, Toby Jia-Jun Li, Nora Bradford, Branda Sun, et al. 2022. Fantastic questions and where to find them: Fairytaleqa–an authentic dataset for narrative comprehension. *arXiv preprint arXiv:2203.13947*.

An Yang, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoyan Huang, Jiandong Jiang, Jianhong Tu, Jianwei Zhang, Jingren Zhou, Junyang Lin, Kai Dang, Kexin Yang, Le Yu, Mei Li, Minmin Sun, Qin Zhu, Rui Men, Tao He, Weijia Xu, Wenbiao Yin, Wenyuan Yu, Xiafei Qiu, Xingzhang Ren, Xinlong Yang, Yong Li, Zhiying Xu, and Zipeng Zhang. 2025. Qwen2.5-1m technical report. *arXiv preprint arXiv:2501.15383*.

Bingsheng Yao, Dakuo Wang, Tongshuang Wu, Zheng Zhang, Toby Jia-Jun Li, Mo Yu, and Ying Xu. 2021. It is ai's turn to ask humans a question: Question-answer pair generation for children's story books. *arXiv preprint arXiv:2109.03423*.

Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, et al. 2024. Justice or prejudice? quantifying biases in llm-as-a-judge. *arXiv preprint arXiv:2410.02736*.

Zhenjie Zhao, Yufang Hou, Dakuo Wang, Mo Yu, Chengzhong Liu, and Xiaojuan Ma. 2022. Educational question generation of children storybooks via question type distribution learning and event-centric summarization. *arXiv preprint arXiv:2203.14187*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.