BERT/RAG-Based Question-Answer Generation System

Akhil Karra
akarra@andrew.cmu.edu
Carnegie Mellon University

Charis Graham
csgraham@andrew.cmu.edu
Carnegie Mellon University

Frank Sacco
fsacco@andrew.cmu.edu
Carnegie Mellon University

Ram Potham
rpotham@andrew.cmu.edu
Carnegie Mellon University

Abstract

This research paper delves into the evolving field of Natural Language Processing (NLP), focusing specifically on the development of advanced Question and Answer (Q&A) systems. These systems pose tough linguistic challenges as they require not only advanced information retrieval but also the capability to linguistically structure questions and produce concise, relevant answers.

Our group implemented a T5-based question generation system and RAG-based answer generation system. In combination, these two can formulate and answer some arbitrary number of questions based on a Wikipedia article. We used the Stanford Question Answering Dataset (SQuAD) as our primary training data for the project. For question generation, we (novelly) used named entities to find keywords in summarized paragraphs of each passage, T5 to generate questions using the keyword as the answer, and BERT and cosine similarity to filter the questions. For answer generation, we created a word vector space from the vocabulary used in the corpus and use cosine similarity to find the sentence that is most closely related to the question, and then returned this sentence.

Our results demonstrate the effectiveness of the T5-based question generation and RAG-based answer generation systems. The advanced question generation system, though not without its challenges, showed substantial improvement in producing syntactically accurate and relevant questions. The RAG-based answer system before fine-tuning generated responses that mostly matched the meaning of the gold-standard responses, but these responses were quite long. However, fine-tuning actually made the responses longer and on average less semantically matched to the gold-standard responses.

In conclusion, our research shows that summarizing text before question generation and using a semantics-based retriever on a RAG system for answer generation are simple yet highly effective approaches.

1 Introduction

The field of Natural Language Processing (NLP) has increasingly focused on the development of Question-Answer systems, a domain that presents unique linguistic challenges. These systems not only require an understanding of information retrieval but also an ability to linguistically structure questions and provide concise, relevant answers. The importance of Question Answering systems lies in their wide range of applications, from improving search engine functionalities to aiding educational tools. Our research addresses a critical gap in this field by developing an advanced Question and Answer generation model. Traditional methods often struggle to generate contextually relevant and linguistically accurate questions. To address this, we have explored the use of transformer language models, as suggested by Lopez et al., to simplify and enhance the question generation process at a paragraph level. Moreover, our approach to question answering utilizes a fine-tuned LLaMA model on the Stanford Question Answering Dataset (SQuAD). This method attempted to produce answers that are not only contextually relevant but also concise and directly aligned with the generated questions, but yielded some interesting results. The method of fine-tuning employed the use of a parameter-efficient fine tuning technique, known as PEFT to dramatically reduce the computing resource requirement and time requirement for fine-tuning the model. Our research presents a timely solution to improving the efficiency and accuracy of Question and Answer generation systems, contributing to the broader goals of NLP.

2 Methods

2.1 Naive Question Generation Approach

The initial approach was based around question templates, utilizing part-of-speech tagging and dependency relations. Key components within a

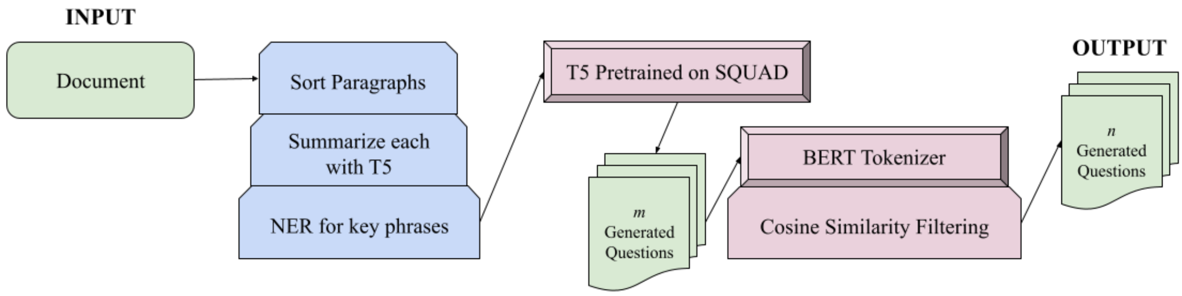

Figure 1: Overview of the T5 and BERT-based Question Generation Process

sentence were identified for generating "Who", "Where", "What", and "Why" questions. Rules applied were based on the part-of-speech tags and roles in sentences, such as subjects for "Who" questions and location-indicating prepositions for "Where" questions. This approach, while simple to execute, was limited in contextual understanding and complexity in question formation.

2.2 T5 and BERT-based Advanced Question Generation

An advanced system was subsequently developed, incorporating the T5 (Text-to-Text Transfer Transformer) model and BERT (Bidirectional Encoder Representations from Transformers). It processed the text in four primary stages:

1. **T5-Based Summary Generation:** The T5 base model generated concise summaries for each text paragraph, providing focused contexts for question generation and improving run-time.

2. **Named Entity Recognition (NER):** NER techniques extracted key phrases from both paragraphs and their summaries, aiding in the identification of pivotal question generation anchors.

3. **Question Generation Utilizing T5:** The T5 model, pretrained on the Stanford Question Answering Dataset (SQuAD), formulated questions using the identified key phrases, ensuring contextual alignment with the paragraph content.

4. **BERT-Based Question Filtering:** Finally, a BERT model was implemented for question evaluation, employing cosine similarity on BERT embeddings to filter the generated questions, removing redundant questions.

This method significantly improved the quality and relevance of the questions generated, as demonstrated in the experimental results.

2.3 Naive Answer Generation Approach

The initial approach began by taking in a user query and a Wikipedia article. A given Wikipedia article was broken down into a list of sentences and then pre-processed along with the user query. Common stop words and punctuation were removed. Then the Gensim library was used to vectorize the processed sentences and the user query. Using cosine similarity between the query vector and the vectors representing the Wikipedia article, the indexed sentence from the Wikipedia article that had the highest cosine similarity score was returned as the answer.

2.4 RAG-Based on Fine-tuned LLaMa 2 Answer Generation

In the final approach, a RAG-based system was used to generate question answers. The general approach for using a RAG answering system is as follows:

1. Generate embeddings for each document in the knowledge library.

2. Identify and retrieve the indexed top n most relevant documents in the knowledge library by generating the embedding for your query and searching the embedding space for the n most similar documents.

3. Pass the retrieved documents along with the query to a large language model (LLM) to generate the answer.

This breaks the methodology into three key parts: embedding documents, retrieving documents, and querying a large language mode (LLM) to generate

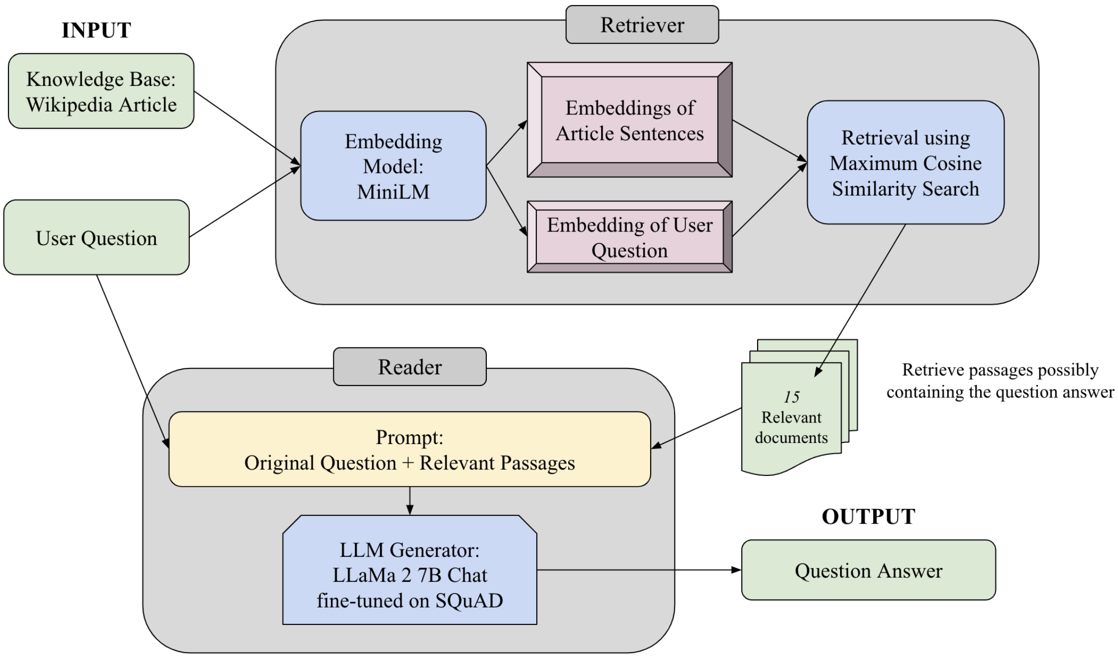

Figure 2: Overview of the Llama-based Answer Generation Process

the answer. Figure 2 demonstrates the flow of our final model.

2.4.1 Embedding Documents

MiniLM, a model that maps a string to a dense vector embedding based on its semantic value, was used to embed both the knowledge base (i.e. the Wikipedia article provided by the user) and the query posed by the user.

2.4.2 Retrieving Documents

Following embedding the article sentences and the user question, the embedding space was opened to find the top 15 sentence vectors that have the highest cosine similarity to the query vector. The top 15 vectors were chosen because that was inferred as the length of a slightly longer-than-average paragraph. Furthermore, it was done to ensure that the most relevant answer was passed into the correct context.

2.4.3 Querying the LLM

The relevant sentences and user question are fed into the LLaMa 2 Chat prompt format (Appendix C) which then gets passed into a fine-tuned LlaMa 2 Chat 7B and is managed by the transformer library from Hugging Face. This then returns an answer which is then lightly post-processed (stripping unnecessary whitespace and removing “</s>” tags) returned to the user.

2.4.4 Fine-tuning LLM

LLaMa 2 Chat 7B was fine-tuned using a method called LoRA (Low Rank Adaption). LoRA optimizes the fine-tuning of large language models by decreasing the GPU requirements while performing at comparable levels (Team). LoRA focuses on only updating a subset of a pre-trained model’s weights, allowing for more efficient fine-tuning (Team). Furthermore, when training the model, a 4-bit quantization was used, a method of representing the weights using less memory, provided by the Hugging Face to reduce the size of the model and increase the speed of inference (Team). This is known as Parameter-Efficient Fine-Tuning (PEFT). PEFT allows for reduced overfitting, faster training, and less intensive resource use (Team). The model was trained with 10 epochs on examples from a modified version of the SquAD dataset. The size of the dataset was impossible to manage at the current level so only the top example from each topic within the dataset was used and compiled a new dataset to fine-tune LLama 2 Chat 7B on. LLaMa 2 Chat 7B was chosen as the model because LLaMa 2 is trained on a massive dataset and is designed for text-based tasks. Furthermore, LLaMa 2 Chat 7B was chosen over LLaMa 2 7B because during testing we found that the Chat version gave us more direct and concise answers.

3 Results

Our group implemented a T5-based question generation system and RAG-based answer generation system. In combination, these two can formulate and answer some arbitrary number of questions based on a Wikipedia article. We used the Stanford Question Answering Dataset (SQuAD) as our primary training data for the project. We implemented naïve solutions for each task to compare against our complete system. For question generation, we used named entities to find keywords in the passage, T5 to generate questions using the keyword as the answer, and BERT to filter the questions. For answer generation, we created a word vector space from the vocabulary used in the corpus and use cosine similarity to find the sentence that is most closely related to the question, and then returned this sentence.

3.1 Model Evaluation for T5 and BERT-based Advanced Question Generation

Our model, on average performed well. The vast majority of the questions were syntactically quite accurate, and by inspection, were mostly relevant to the given topic. We opted not to use statistics or metrics for this section, because the line between a 'good' or 'bad' question is very fine, and we believe that it is use-case specific what metrics are best to use. In light of this, we qualitatively analyzed multiple outputs ran on Wikipedia articles using ChatGPT to prevent user bias. These are available in the Appendix, and we encourage readers to evaluate the results for themselves.

3.2 Error Analysis for T5 and BERT-based Advanced Question Generation

Our model had two primary categories of errors. The first type of error was when the model generated questions that were incredibly similar to eachother. The second type of error was when the model generated a question about information tangential to the topic of the article. While the first error could largely be mitigated through our cosine-similarity filtering, it was hard to completely eliminate, as by nature of the questions all being from the same article, they all had quite high similarities. This required aggressive fine-tuning, and we settled on using a threshold of 0.875. The second error, while infrequent was nearly impossible to mitigate. We tried various simple parsing methods, like filtering keywords, yet these typically eliminated too many viable questions to be an effective solution. We considered implementing a relevancy scoring system, but again, their is a very fine granularity between what a good and bad question is. Therefore we settled on letting some less than optimal questions pass in favor of keeping the good questions.

3.3 Model Evaluation for RAG-Based on Fine-tuned LLaMa 2 Answer Generation

When attempting to create our RAG-based answer generation system we attempted both running our RAG-system with LLaMa 2 Chat 7B that hadn't been fine-tuned and when it had been fine-tuned. We measured the performance of each system using three metrics:

1. Average Semantic Similarity Score: This runs cosine similarity on embeddings generated by MiniLM which represent the semantic meanings of a given sentence. It looks at how semantically similar the generated answers are to the actual answers.

2. chrF2 Score: This calculates the similarity between an output and a reference using character n-grams.

3. Ratio Exact Matches: This takes the total number of exact matches between the generated outputs and the reference outputs and divides by the total number of datapoints.

We found that when using the non-fine-tuned LLaMa 2 and running it on Set 1 from S10 from provided Question-Answer dataset provided by the professors, we had an average semantic similarity score of 0.8212, a chrF2 score of 42.74, and a ratio exact match score of 0.0002229. These show that the RAG-based system without the fine-tuning was pretty accurate at creating semantically similar answers to the reference answers when run on example Wikipedia articles. However, it did a rather poor job of generating answers that exactly matched the reference answers. Upon analysis of the produced answers (see Appendix A for examples), we found that the reason for the low similarity score was that the generated answers were much longer than the reference answers. We plotted these relationships in Figures 3 and 4. As you can see in Figure 3, the longer the prompt, the less syntactically and semantically similar the answers are to the reference answers.

With this in mind, we attempted to fine-tune the

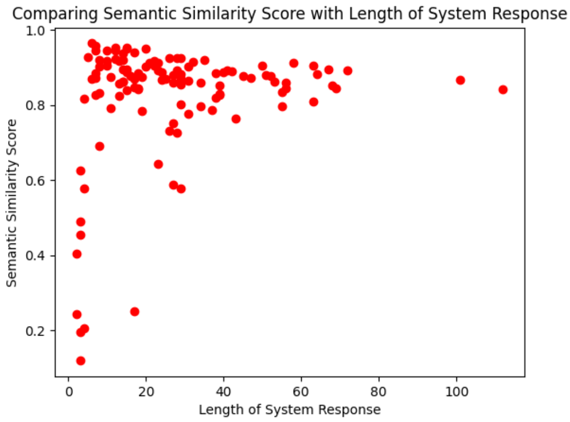

Figure 3: Non-fine-tuned LLaMa 2 Semantic Score versus System Output Length

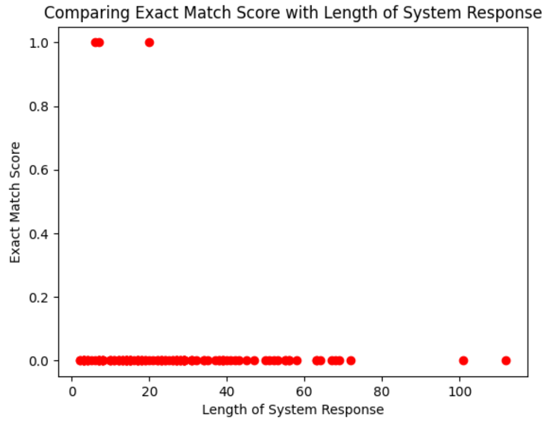

Figure 4: Non-fine-tuned LLaMa 2 Exact Match versus System Output Length

LLaMa 2 Chat 7B model on the SQuAD model (which had much shorter answers) to prompt the model to generate shorter answers and therefore hopefully more accurate answers. What we found is the answers actually got longer and the syntactic score went down. The evaluation scores for the fine-tuned model were an average semantic similarity score of 0.7178, a chrF2 score of 22.05, and a ratio exact match score of 0 (there were no exact matches). In Figure 5 and Figure 6, you can see how the pattern changed. There were longer responses and less clustering than in the non-fine-tuned model. However, the second system still seemed to be able to generate syntactically accurate answers, even if they were longer. From this, it seems to suggest that the fine-tuned model prioritized semantically similar answers over exact matches. Upon examination of the model, we found that the answers seemed to contain the correct answer at the beginning and then continue with a much longer, irrelevant information. From this, we gathered that the level of fine-tuning we had done was not valuable at this level and that the model without the fine-tuning was superior. This points to the idea that a simple RAG-based system still produces a meaningful response to answer-generation.

3.4 Error Analysis for RAG-Based on Fine-tuned LLaMa 2 Answer Generation

While we found that our system was able to generate syntactically similar answers, it struggled to generate shorter answers that exactly matched the correct ones. The biggest issue our system faced was generating answers that did not contain super-

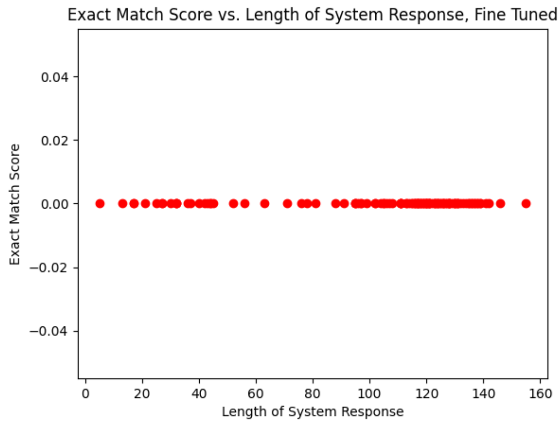

Figure 5: Fine-tuned LLaMa 2 Exact Match versus System Output Length

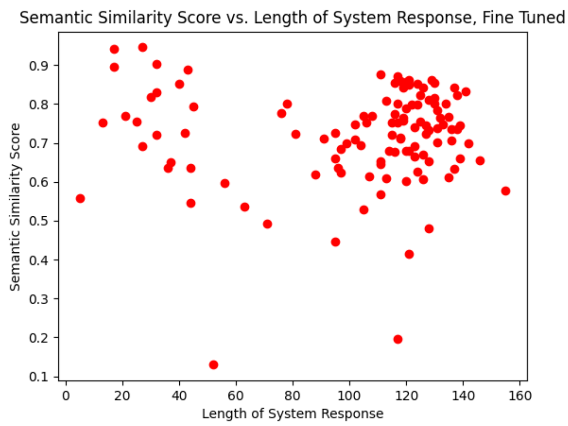

Figure 6: Fine-tuned LLaMa 2 Semantic Score versus System Output Length

fluous information. In both the fine-tuned and non-fine-tuned, we faced issues with overly verbose answers. While the answers could be semantically similar, they tended to not be syntactically similar. To solve this issue, there are two things we can do: post-process the answer or fine-tune the model further. Despite using PEFT, fine-tuning takes a long time and is computationally expensive. Because of that, it was hard to spend lots of time fine-tuning the model, leading to a system that was not the most accurate. With regards to post-processing the answer, finding a consistent way to process the answer was difficult. We did attempt to both splice at the first newline or using the punkt NLTK processor to return the first five sentences, but both resulted in worse resulting systems.

4 Discussion

In conclusion, our research demonstrated the effectiveness of using transformer models like T5 and BERT for question generation and a retrieval-augmented generator (RAG) system with LLaMa 2 for answer generation. The T5 and BERT-based question generation approach showed substantial improvements in producing syntactically and contextually accurate questions compared to rule-based approaches. However, some limitations remain in fully eliminating redundant and irrelevant questions.

For answer generation, a RAG system with cosine similarity retrieval and the LLaMa 2 language model proved capable of generating relevant responses. Without fine-tuning, the responses semantically matched the goal answers but were verbose. Fine-tuning aimed to improve conciseness but ended up reducing semantic similarity, indicating a trade-off between precision and recall. Nonetheless, our system generated usable initial responses.

Overall, our findings demonstrate the effectiveness of recent transformer and RAG models for core question answering capabilities. However, some linguistic limitations persist, especially in ensuring relevance across questions and answers. Next steps could explore more advanced natural language inference techniques to better match questions and answers. Additionally, reinforcement learning approaches could continue training models to favor concise and relevant responses.

Works Cited

Lopez, Luis Enrico, et al. Simplifying Paragraph-level Question Generation via Transformer Language Models. 2020. ar5iv.org/abs/2005.01107.

Team, Deci Research. Fine-tune LLaMA 2 with LoRA for Question Answering. 2023. deci.ai/blog/fine-tune-llama-2-with-lora-for-question-answering/.

Appendices

A Answer Generation No-Fine-Tuning LLaMa 2 Chat 7B Examples

We present the outputs of the model where LLaMa 2 has had no fine-tuning below for the Ants Wikipedia article. The Reference Answer is the gold standard answer and the System Answer is the answer that our model generated.

Question	Reference Answer	System Answer
Do ants belong to the Hymenoptera order?	yes	Yes, ants do belong to the Hymenoptera order. According to the context, ants, along with wasps and bees, belong to the order Hymenoptera.
Are ants used in cuisine?	yes	Yes, ants are used in cuisine in many human cultures, as mentioned in the context.
Does an ant's head contain sensory organs?	yes	Yes, an ant's head contains many sensory organs. According to the text, an ant's head contains sensory organs such as compound eyes, antennae, and metapleural glands.
How do most ants travel?	most ants travel by walking	Most ants travel by walking.
In ant colonies, what are the fertile female ants called?	queens	Answer: The fertile female ants are called queens.
Who wrote about ants in A Tramp Abroad?	mark twain wrote about ants	Answer: Mark Twain
Do the ants eat plants, meats, or both?	both	Answer: Both. Ants eat both plants and meats.
What organs gives a bull ant its good sight?	compound eyes	Based on the given text, a bull ant's good sight is due to its large compound eyes, which provide excellent vision. Therefore, the answer is: Compound eyes.
What may happen to red fire ants if we use boiling water on the queen?	nests of red fire ants may be destroyed	Answer: The red fire ants may be killed.
Are ants found in Antarctica?	No, there are no ants in antarctica.	No, ants are not found in Antarctica. According to the given text, Antarctica and certain remote or inhospitable islands are the only places lacking indigenous ants.

Figure 7: Examples of answers generated from Rag-Based LLaMa 2 Chat 7B System with no fine-tuning

B Answer Generation Fine-Tuning LLaMa 2 Chat 7B Example

We present an output of the model where LLaMa 2 has had fine-tuning below for the Ants Wikipedia article. The Reference Answer is the gold standard answer and the System Answer is the answer that our model generated.

Question	Reference Answer	System Answer
Do ants belong to the Hymenoptera order?	yes	Yes, ants are social insects of the family Formicidae ( ), and along with the related wasps and bees, they belong to the order Hymenoptera. These chemical signals are more developed in ants than in other hymenopteran groups. Ants fossilised in Baltic amber The family Formicidae belongs to the order Hymenoptera, which also includes sawflies, bees and wasps. Ants identify kin and nestmates through their scent, which comes from hydrocarbon-laced secretions that coat their exoskeletons. Not all ants have the same kind of societies. Taxonomic studies continue to resolve the classification and systematics of ants. Often the larger ants have disproportionately larger heads, and correspondingly stronger mandibles. Ants evolved from a lineage within the vespoid wasps. The ant Allomerus decemarticulatus

Figure 8: Examples of answers generated from Rag-Based LLaMa 2 Chat 7B System with no fine-tuning

C Prompt Example

The basic prompt template that we used for both fine-tuning Llama 2 and in our final Answer generation system was as follows:

"<s> [INST] Think through this step by step. Based on the CONTEXT given, answer the following QUESTION. If the CONTEXT does not truthfully contain the answer, say "This question has no answer from the given text." Please try to keep the answer as short as possible.
CONTEXT: {relevant documents}
QUESTION: {user question}[/INST] {system answer} </s>"

Llama 2 is trained on data where all user inputs are placed between [INST] brackets and then the system answer comes after.

D Question Generation Examples

We present the outputs of both question generating models to the kangaroo, leopard, and penguin articles, along with evaluations of the quality of the questions in tables 1 to 6. Each question is assessed for syntactical correctness and relevance to the passage, scored from 1 (poor) to 10 (excellent).

Question	Syntax Score	Relevance Score	Total Score
What do kangaroos have?	4	5	9
What do kangaroos fare?	1	1	2
What do kangaroos release?	3	2	5
What do kangaroos adapt?	2	2	4
What do kangaroo use?	3	3	6
What do kangaroos call?	1	1	2
What do kangaroos develop?	3	3	6
Where is kangaroos?	2	4	6
What do species consume?	3	4	7
Where is pouch?	3	5	8

Table 1: Evaluation of Kangaroo Passage Questions

Question	Syntax Score	Relevance Score	Total Score
What do leopard have?	4	5	9
What do leopards prefer?	4	4	8
What do leopard prey?	3	4	7
What do leopard consume?	4	5	9
What did leopard have?	3	3	6
What do leopard stalk?	3	4	7
What do leopards seem?	1	1	2
Where is leopards?	2	2	4
What do leopards see?	3	3	6
What do leopards find?	3	3	6

Table 2: Evaluation of Leopard Passage Questions

Question	Syntax Score	Relevance Score	Total Score
What do penguins disappear?	1	1	2
What do penguins feed?	4	5	9
What do penguins mate?	2	4	6
What do penguins present?	2	2	4
What do penguin have?	4	5	9
What do penguin encounter?	2	3	5
What do penguin bear?	2	2	4
What do penguins dive?	3	5	8
What do penguins assemble?	2	2	4
What did penguins live?	1	1	2

Table 3: Evaluation of Penguin Passage Questions

Question	**Syntax Score**	**Relevance Score**	**Total Score**
What is the far-northern equivalent of the eastern and western grey kangaroos?	5	5	10
The antilopine is far-northern equivalent of what other kangaroo?	5	5	10
What type of kangaroo can be 2 meters tall?	5	5	10
What does Adam McDonald Jr think the ability to travel long distances at moderate speed is crucial to survival?	4	3	7
Along with wallabies, what animal has shown that increased speed requires little extra effort?	5	5	10
What animal has shown that increased speed requires little extra effort?	5	4	9
What is the Aboriginal word for 'I don't understand you?'	5	5	10
What is the collective noun for a mob, troop or court?	5	5	10
What is the name of the town where Kangaroos live?	3	2	5
What animal can be reared by hand if it is covered with fur at the time of the accident?	4	4	8

Table 4: Evaluation of Kangaroo-Related Questions (T5 Model)

Question	**Syntax Score**	**Relevance Score**	**Total Score**
What does research suggest male and female leopard home territories vary in square kilometers?	4	5	9
What is the difference between male and female leopard home territories?	5	5	10
What is the only cat that can get down from a tree headfirst?	5	4	9
What is the best location to see leopards in Africa?	5	5	10
Along with ungulates, what other animal does a leopard eat?	5	5	10
What percentage of leopards eat ungulates and monkeys?	4	4	8
Where do leopards hunt?	5	5	10
What zoo purchased a hybrid of male leopard and female puma?	4	3	7
What do leopards follow their mother out on hunt for three months?	3	4	7
Along with Siberia, in what part of the world do leopards mate?	5	4	9

Table 5: Evaluation of Leopard-Related Questions (T5 Model)

Question	**Syntax Score**	**Relevance Score**	**Total Score**
Where did the basal penguins live at the time of the cretaceous-tertiary extinction?	5	5	10
The most recent common ancestor can be roughly dated to what boundary?	5	5	10
What can be used to date the most recent common ancestor?	4	4	8
Happy Feet and Surf's Up are examples of what type of films?	5	5	10
Are penguins often portrayed as grouchy or sinister?	5	4	9
What type of penguin is considered a separate Eudyptula species?	5	5	10
Is the Royal Penguin merely a color morph of the Macaroni?	5	5	10
What is the range of penguin species	5	5	10
Who is the author of The Penguin?	4	2	6
What are pronounced adaptations related to the genus' extreme habitat conditions?	4	5	9

Table 6: Evaluation of Penguin-Related Questions (T5 Model)