ChatPI Report

Andrei Cozma
Graduate Student
University of Tennessee
acozma@vols.utk.edu

Manan Patel
Undergraduate Student
University of Tennessee
mpatel65@vols.utk.edu

Tulsi Tailor
Undergraduate Student
University of Tennessee
ttailor@vols.utk.edu

Zac Perry
Undergraduate Student
University of Tennessee
zperry4@vols.utk.edu

I. Introduction

In this report, we explore constructing and evaluating key components traditionally used in building chatbots, focusing on their application in a multilingual context. Our project centers around "The Sign of the Four" by Sir Arthur Conan Doyle, sourced from Project Gutenberg eBooks. The main objective is to assess various natural language processing (NLP) pipelines for question answering, translation, and summarization using pre-trained models from Hugging Face. We independently evaluate each pipeline through various metrics and integrate them into a simple chat interface. This approach allows us to understand the effectiveness and limitations of each NLP component in isolation and as part of a cohesive chatbot system.

II. Methodology

Our methodology is structured into three sections, each dedicated to a key pipeline in NLP and chatbot technology. These sections encompass the core functionalities of chatbots: understanding and responding to user queries, communicating in multiple languages, and condensing large text into coherent summaries. Below, we explore our methodology for each pipeline in detail.

A. Question-Answering

Question-answering (QA) in NLP involves a model generating or extracting answers from provided text [2]. We aim to evaluate QA capabilities in our project, focusing on "The Sign of the Four" by Sir Arthur Conan Doyle.

We selected five 300-word excerpts from the novel, encompassing key aspects like character identities, crime scene details, evidence, and the case against the perpetrator. We developed a quiz for each section with nine specific questions, each designed for concise answers (one to four words) found within the text. This approach intends to benchmark the models' performance in accurately identifying relevant information.

We chose various models from Hugging Face for this assessment, representing different QA methodologies: - DistilBERT Models: *distilbert-base-cased-distilled-squad* and *distilbert-base-uncased-distilled-squad*, balancing performance and efficiency [3,4]. - RoBERTa Models: *deepset/roberta-base-squad2* and *deepset/roberta-large-squad2*, fine-tuned for QA and adaptable to unanswerable questions [5,6]. - DeBERTa Models: *deepset/deberta-v3-base-squad2* and *deepset/deberta-v3-large-squad2*, using advanced techniques for QA enhancement [7,8,9]. - ELECTRA Model: *deepset/electra-base-squad2*, featuring a generator-discriminator architecture for NLP tasks [10,11,12].

Our evaluation will compare the model responses with the expected answers, using various metrics to assess accuracy and relevance. This analysis aims to identify each model's strengths and weaknesses in the context of our project.

B. Translation

In our project's translation pipeline, we focus on enabling the chatbot to handle translations between English and French. This aspect is crucial for assessing the performance of machine translation models, which are vital in NLP for converting text between languages using sophisticated algorithms [13].

For this task, we use the same 300-word sections from "The Sign of the Four" and the corresponding answers from our question-answering quiz. Our selected translation models are specifically chosen to evaluate their effectiveness in a bilingual context:

- Opus Models: We employed two distinct models from the Opus series: *Helsinki-NLP/opus-mt-en-fr* for translating from English to French and *Helsinki-NLP/opus-mt-fr-en* for the reverse translation. These models are exciting as they are each fine-tuned for specific language pairs, providing a focused approach to translation between English and French [14,15].

- Multilingual Model: *facebook/m2m100_418M*, a comprehensive model trained on a diverse Many-to-Many dataset, covering 100 languages, including English and French. Its training on a multilingual dataset suggests a broader, more versatile translation capability compared to the Opus models, which are more specialized [16,17,18].

A crucial part of our evaluation will involve observing how these models perform in a cyclical translation process – translating English answers to French and then back to English. This will not only test their accuracy in maintaining the essence of the original text but also their effectiveness in handling nuances of language in a to-and-fro translation cycle.

It will be particularly interesting to compare the performance of the specialized Opus models against the more generalized capabilities of the *facebook/m2m100_418M* model. Such a comparison will provide insights into the strengths

and limitations of using specialized versus multilingual models in practical translation tasks. The metrics used for this evaluation, along with the results, will be discussed in the later sections of the report, providing a comprehensive overview of the effectiveness of these models in a multilingual communication context.

C. Summarization

The summarization pipeline in our project is geared towards condensing key sections of the novel into coherent summaries. This pipeline tests the models' ability to capture and articulate the essential details and themes from the text [19]. For this purpose, we utilized three different models from the Hugging Face summarization pipeline, ensuring a comparative analysis of their performance.

We continued using the same 300-word sections selected for the protagonist, antagonist, crime, evidence, and resolution. These sections served as the basis for our summarization task, aiming to provide concise and contextually accurate summaries. The chosen model for this task was:

- DistilBART (*distilbart-cnn-12-6*): This model, known for its efficiency and reduced size, is the default choice for the summarization pipeline. It was trained on news articles, making it suitable for summarizing shorter texts like our 300-word excerpts [20,21,22]. - BART Summarization Model (*bart_summarisation*): Specifically designed for summarization, this model was trained on dialogue datasets, which makes it potentially effective for sections of the novel containing dialogue [20,23,24]. - PEGASUS-X Large Book Summary (*pegasus-x-large-book-summary*): A model trained on various texts, including novels. Its design for long-form narrative summarization made it an intriguing choice for our shorter excerpts, providing a test of its adaptability to different text lengths [25,26,27].

Each model's performance was evaluated based on how well they summarized the text, maintaining the context and key information. The comparison of these models, trained on different datasets, aimed to shed light on which model is more adept at summarizing shorter pieces of text like our selected novel excerpts. This evaluation was integral to understanding the capabilities and limitations of different summarization models in the context of chatbot technology.

D. Evaluation

The evaluation of each NLP pipeline in our project plays a pivotal role in understanding the strengths and weaknesses of the models we've employed. By using a range of metrics, we aimed to assess the models' performance in question-answering, translation, and summarization tasks, ensuring a thorough analysis of their capabilities. By integrating these metrics into our evaluation process, we gained valuable insights into the behavior and accuracy of our models.

1) Cosine similarity score using Spacy: This metric, leveraging the Spacy library with the *en_core_web_lg* model, measures semantic similarity by averaging word vectors. While it offers a basic assessment of semantic alignment, especially after removing stop-words and punctuation for clarity, it's somewhat limited in depth, leading us to consider more sophisticated evaluation tools.

2) BERTScore: For a more detailed textual analysis, we turned to BERTScore, using the *microsoft/deberta-xlarge-mnli* model. It provides a nuanced evaluation by comparing token-level similarities between candidate and reference sentences. This approach, known for its strong correlation with human judgment, is particularly effective for our project's diverse NLP tasks. The model's high ranking in the "WMT16 To-English Pearson Correlation" further emphasizes its ability to align closely with human quality assessments.

3) Rouge Score: Rouge Score was applied across all our pipelines, with a historical focus on summarization and translation. Comprising metrics like Rouge-1, Rouge-2, Rouge-L, and sum, it evaluates the overlap of n-grams and longest common sub-sequences, offering a comprehensive view of how well our models capture and reproduce key elements of the original text.

III. Results and Analysis

A. Question-Answering

We comprehensively evaluated the question-answering pipeline using various models, including DistilBERT, RoBERTa, DeBERTa, and ELECTRA. We used several metrics to test our question-answering pipeline, such as Bert, Rouge, and Spacy's cosine similarity score.

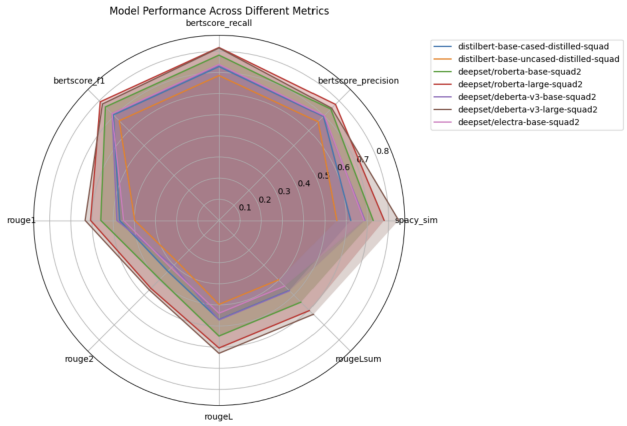

Fig. 1. Q/A: Radar plot of comparing model performance across different metrics

The radar plot shown in Figure 1 provides a visualization of how well each model performs using a variety of metrics. In particular, *distillbert-base-uncased-distilled-squad* generally performs poorly, and *deepset/deberta-v3-large-squad2* performs very well overall. DeBERTa is known for its effectiveness in processing complex language structures and understanding context due to its large-scale training on the Squad2 dataset. We can see that Bert is the most prevalent metric to evaluate our question-answering pipeline due to its ability to measure semantic similarity using contextual embeddings, perform token-level assessment accommodating synonyms, handle variations in sentence length, and simplify

the evaluation process without needing reference answer tokenization.

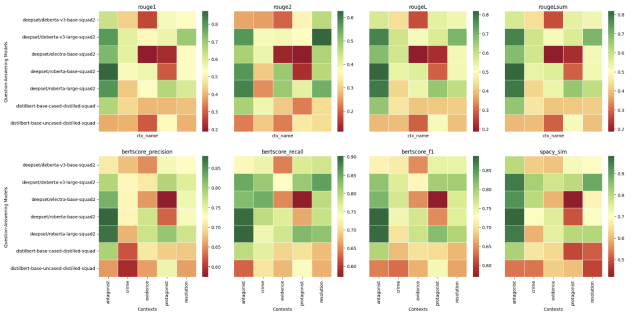

Fig. 2. Q/A: Avg. metric scores for selected contexts across the two models

To explain the performance variations across different sections and questions deeper, we generated a heat-map, as shown in Figure 2, that breaks down the scores for different contexts. The x-axis represents other models from each context, such as protagonist, antagonist, crime, evidence, and resolution. The y-axis represents the model names and the color intensity, with red being the lowest and green being the highest.

The heatmap offers insights into the models' performance on specific sections or questions. For instance, in the evidence section, question 3 poses challenges for several models, as lower scores indicate. This could suggest that specific nuances or complexities in the text are more challenging for the models to capture accurately. Also, the antagonist section exhibits a relatively high variance in score. The observed differences in performance could be attributed to factors such as model size, training dataset, or architectural variations.

B. Translation

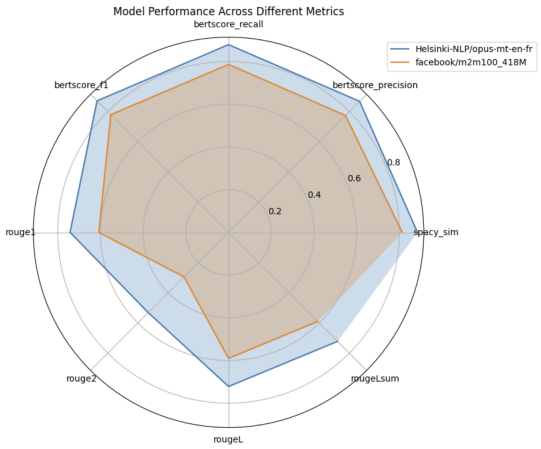

Fig. 3. Translation: Radar plot of comparing model performance across different metrics

In Figure 3, *Helsinki-NLP/opus* outperforms Facebook's *m2m100_418M* model in our evaluations, primarily due to Opus being trained explicitly for English-to-French and French-to-English translation tasks. In contrast, the *m2m100_418M* model is designed to be multilingual, allowing it to handle translations across various languages. However, this multilingual capability comes at the cost of specificity, making Opus a more suitable choice for tasks specifically focused on English and French translation.

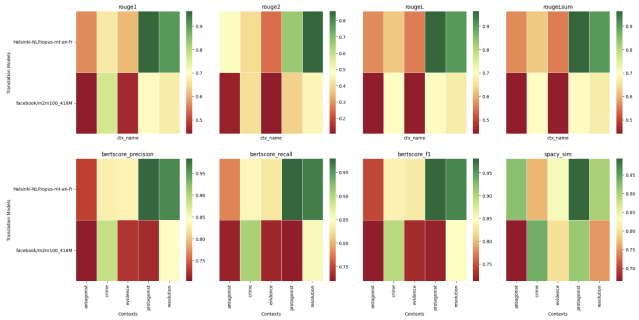

Fig. 4. Translation: Avg. metric scores for selected contexts across the two models

This plot, in Figure 4, illustrates the average scores for selected contexts across two models, Opus and *m2m100_418M*. In the heat-map, darker colors signify higher similarity scores, while lighter colors indicate lower scores. The x-axis displays the correct answers for each question within the context, and the y-axis represents the two models being compared.

In the evaluation of translation models, two prominent models, namely the *Helsinki-NLP/opus-mt-fr-en* and the multilingual model (*facebook/m2m100_418M*), revealed intriguing patterns and discrepancies. Notably, the *Helsinki-NLP/opus-mt-fr-en* model exhibited an occasional looping behavior during the translation from French to English. In specific instances, the model would predict the same token(s), such as "right leg" or "skin," repeatedly until it reached the maximum number of completion tokens. This unique behavior introduces a potential concern regarding the model's robustness in handling certain input patterns and may impact the overall translation quality.

A distinct difference emerged in the translation of names, exemplified by "Jonathan Small." While the *Helsinki-NLP/opus-mt-fr-en* model consistently provided the correct translation, the multilingual model occasionally produced unexpected results, translating "Jonathan Small" back to English as "Jonathan Little." This discrepancy raises questions about the multilingual model's handling of specific entities and underscores the need for further investigation into the nuances of name translation.

Moreover, the analysis revealed a general trend with both models. As the input sequence length increased, semantic information was lost during the repeated or cyclical translation process. This observation underscores the challenge of maintaining context and meaning in longer input sequences, prompting considerations for optimizing translation models for diverse input lengths.

In examining specific instances, anomalies were identified in the translation output of the multilingual model. For

example, when the input sequence was "Jonathan Small," the model predicted an unexpected translation, such as "by Sherlock Holmes," during the back-translation to English. Although the impact on similarity scores was minimal due to removing stop-words, a significant penalty was observed in the BERTScore (where stop-words were retained) for including the additional token. These anomalies emphasize the importance of a nuanced evaluation, considering different evaluation metrics and the specific characteristics of the translated content.

Additionally, The multilingual model tended to add more extra prepositions, such as "by," and articles, like "the," in the translated output back to English compared to the Opus model. This discrepancy could result in how the correct answer was supposed to be answered.

C. Summarization

Our evaluation of the summarization pipeline involved three distinct models: DistilBART, BART_summarisation, and PEGASUS-X.

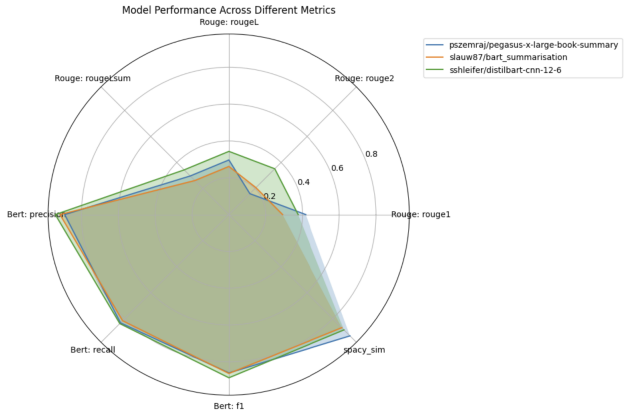

Fig. 5. Summarization: Radar plot of comparing model performance across different metrics

The radar plot above represents the performance of the summarization models across all of the metrics we used. This provides a clear visual representation and directly compares each model and its associated scores.

When further examining the plot, we first see that all three models have relatively higher scores for each BERT metric and the Spacy_sim metric when compared to ROUGE. This could be because ROUGE evaluates generated summarizations by counting the number of overlapping units and word sequences [28]. Since the resulting text will be a summarization, our models may have used only a few overlapping units and word sequences from the original text to produce this summarization, causing it to score lower. On the other hand, the BERT scores represent the similarity score for each token in the candidate sentence with each token in the reference sentence [29]. Comparing the similarity of each token in each text could have produced a higher similarity overall, resulting in a much higher score.

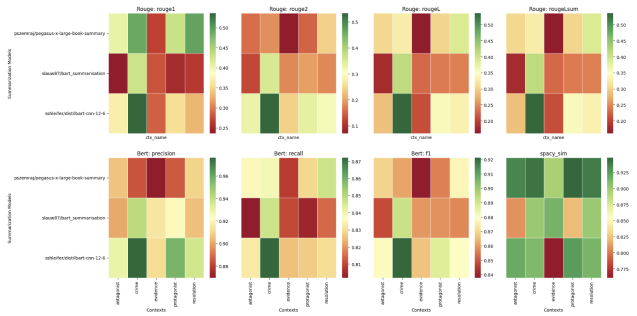

Fig. 6. Summarization: Avg. metric scores for selected contexts across the two models

The x-axis represents different contexts, such as antagonist, crime, evidence, protagonist, and resolution. In contrast, the y-axis represents the different model names, with the color intensity of red being the lowest and green indicating the highest scores.

These metrics show that Spacy_sim yields the most accurate results overall. Spacy_sim performs much better than other metrics because while comparing the metrics, we removed stop words and punctuation to yield more accurate results. One exciting outcome was how well the DistilBART model performed on each metric for the crime context. This could be because this model was trained on various news articles, some of which could have been about crime and could have aided the model in better generalization and summarization. Additionally, all models performed relatively well for the crime context compared to the other contexts for each metric. This consistency could be due to the excerpt providing sufficient context to generalize and summarize the text well. Another interesting trend was how the models often performed poorly for the evidence context across each metric. The cause could be that the excerpt provided lacked specific context surrounding the evidence, causing the model to summarize the text poorly and result in very low scores.

When comparing the metrics between these three models, the DistilBART model consistently performed well for each metric. Even though it did not perform well for the evidence context for the Spacy_sim metric, it still had excellent results across all other contexts and metrics.

IV. Conclusion

This report comprehensively evaluated key chatbot components in a multilingual context centered around "The Sign of the Four" by Sir Arthur Conan Doyle. Our exploration encompassed diverse NLP pipelines, specifically question-answering, translation, and summarization, utilizing pre-trained models from Hugging Face. Through rigorous testing and analysis, we have identified the strengths and limitations of each pipeline, both as standalone functionalities and as integrated elements of a chatbot system. This investigation provides valuable insights into the efficacy of these NLP components in a chatbot interface and underscores the complexities and challenges of applying these technologies in multilingual contexts.

REFERENCES

[1] Doyle, A. C. (2000). The Sign of the Four. Project Gutenberg. https://www.gutenberg.org/ebooks/2097

[2] Huggingface.co, 2023. https://huggingface.co/docs/transformers/tasks/question (accessed Dec. 04, 2023)

[3] "distilbert-base-uncased-distilled-squad · Hugging Face," huggingface.co, Apr. 05, 2023. https://huggingface.co/distilbert-base-uncased-distilled-squad (accessed Dec. 04, 2023)

[4] "squad · Datasets at Hugging Face," huggingface.co, Apr. 12, 2023. https://huggingface.co/datasets/squad

[5] "roberta-base · Hugging Face," huggingface.co. https://huggingface.co/roberta-base

[6] "squad_v2 · Datasets at Hugging Face," huggingface.co. https://huggingface.co/datasets/squad_v2

[7] "microsoft/deberta-v3-base · Hugging Face," huggingface.co. https://huggingface.co/microsoft/deberta-v3-base

[8] "deepset/deberta-v3-base-squad2 · Hugging Face," huggingface.co, Apr. 05, 2023. https://huggingface.co/deepset/deberta-v3-base-squad2 (accessed Dec. 04, 2023)

[9] "deepset/deberta-v3-large-squad2 · Hugging Face," huggingface.co. https://huggingface.co/deepset/deberta-v3-large-squad2

[10] "ELECTRA," huggingface.co. https://huggingface.co/docs/transformers/model_doc/electra

[11] K. Clark, M.-T. Luong, G. Brain, Q. Le Google Brain, and C. Manning, "ELECTRA: PRE-TRAINING TEXT ENCODERS AS DISCRIMINATORS RATHER THAN GENERATORS." Available: https://openreview.net/pdf?id=r1xMH1BtvB

[12] "deepset/electra-base-squad2 · Hugging Face," huggingface.co, Apr. 05, 2023. https://huggingface.co/deepset/electra-base-squad2 (accessed Dec. 04, 2023)

[13] "Translation," huggingface.co. https://huggingface.co/docs/transformers/tasks/translation

[14] "Helsinki-NLP/opus-mt-en-fr · Hugging Face," huggingface.co. https://huggingface.co/Helsinki-NLP/opus-mt-en-fr (accessed Dec. 04, 2023)

[15] "OPUS - an open source parallel corpus," opus.nlpl.eu. https://opus.nlpl.eu/

[16] "facebook/m2m100_418M · Hugging Face," huggingface.co. https://huggingface.co/facebook/m2m100_418M

[17] A. Fan et al., "Beyond English-Centric Multilingual Machine Translation." Accessed: Dec. 04, 2023. [Online]. Available: https://arxiv.org/pdf/2010.11125.pdf

[18] S. Jiang, "Transformer Align Model," Shaojie Jiang's Homepage, May 16, 2020. https://shaojiejiang.github.io/post/en/transformer-align-model/ (accessed Dec. 04, 2023)

[19] "Summarization - Hugging Face NLP Course," huggingface.co. https://huggingface.co/learn/nlp-course/chapter7/5?fw=pt

[20] "BART," huggingface.co. https://huggingface.co/docs/transformers/model_doc/bart

[21] "sshleifer/distilbart-cnn-12-6 · Hugging Face," huggingface.co. https://huggingface.co/sshleifer/distilbart-cnn-12-6

[22] "cnn_dailymail · Datasets at Hugging Face," huggingface.co. https://huggingface.co/datasets/cnn_dailymail

[23] "slauw87/bart_summarisation · Hugging Face," huggingface.co, Dec. 27, 2022. https://huggingface.co/slauw87/bart_summarisation

[24] "samsum · Datasets at Hugging Face," huggingface.co. https://huggingface.co/datasets/samsum

[25] "PEGASUS-X," huggingface.co. https://huggingface.co/docs/transformers/model_doc/pegasus_x (accessed Dec. 03, 2023)

[26] "pszemraj/pegasus-x-large-book-summary · Hugging Face," huggingface.co, Nov. 30, 2022. https://huggingface.co/pszemraj/pegasus-x-large-book-summary (accessed Dec. 03, 2023)

[27] "kmfoda/booksum · Datasets at Hugging Face," huggingface.co, Nov. 28, 2023. https://huggingface.co/datasets/kmfoda/booksum (accessed Dec. 03, 2023)

[28] C.-Y. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries." Available: https://aclanthology.org/W04-1013.pdf

[29] "BERT Score - a Hugging Face Space by evaluate-metric," huggingface.co. https://huggingface.co/spaces/evaluate-metric/bertscore