MistralAI-XRay: Enhanced Diagnostic Report Generation with VLM and RAG

Rahul Kumar
Northeastern University
Boston, USA
kumar.rahul4@northeastern.edu

Avnish Patel
Northeastern University
Boston, USA
patel.avni@northeastern.edu

***Abstract*—This project uses a hybrid artificial intelligence framework that combines the Vision Language Model - ALBEF (Align Before Fuse) with the Retrieval-Augmented Generation (RAG) framework to automatically generate medical reports from Chest X-ray images. Initially trained on the Indiana University Open-i dataset, ALBEF retrieves relevant medical reports as contextual references. These reports are then used as prompts in the RAG framework, which works with the open-source generative model Mistral AI open-mistral-7b, to create detailed and accurate radiological findings. This approach generates reports that are coherent with a radiologist's written report and can also be tailored to specific clinical settings through prompt engineering. The goal is to enhance report accuracy and reduce the time healthcare professionals spend on the time-consuming task of report generation.**
Github Repo: MistralAI-XRay Code
Dataset: Open-i

I. Introduction

The need for quick and accurate medical report generation in radiology is critical for timely and efficient patient care [1]. Traditional methods, which rely on manual report writing by radiologists, are time-consuming and prone to human error. Earlier methods for automatic report generation treated this task as image captioning or a generative task. However, these methods often produced reports with self-contradictory claims and lacked proper clinical language [2]. Another approach treats report generation as a retrieval-based task, where the model retrieves the most relevant reports for a given X-ray image from a corpus of medical reports [3]. This method ensures coherence with expert-written reports and, because the medical terminology used in diagnoses is relatively limited, retrieving from a large corpus can provide accurate reports for most diagnoses. Retrieval-based methods have been shown to achieve higher accuracy than current generative methods [4]. This work aims to push the accuracy further for more precise and efficient diagnoses.

However, purely retrieval-based models [2] [3] have several limitations that can hinder their effectiveness. They often include irrelevant information, particularly in cases where no findings are present, and can generate noise from prior reports. Additionally, these models may suffer from duplicate content and can produce incoherent information by combining sentences from different patients' reports, which can lead to hallucinations and inconsistent outcomes. On the other hand, the rise of powerful open-source generative Large Language Models (LLM) such as Mistral AI [7] presents a promising alternative. This model can generate relevant content based on prompts, offering a more tailored approach. However, it also has its drawbacks, particularly in lacking up-to-date and domain-specific information that is crucial in medical contexts. To address these challenges, the integration of retrieval-augmented generation (RAG) emerges as a compelling solution [5]. RAG combines the strengths of both retrieval-based and generative approaches, enhancing the factual grounding of generated reports while reducing the risk of hallucinations. By leveraging this hybrid method, we can significantly improve the quality and accuracy of medical reports, making them more reliable and accurate.

This project leverages Mistral AI to create concise chest X-ray reports using the Retrieval-Augmented Generation (RAG) framework. The process begins by employing the ALBEF model to search and retrieve relevant reports from the Indiana University Open-i dataset. These retrieved reports serve as prompts within the RAG framework, which Mistral's open-mistral-7b model uses. This combination allows for the synthesis of detailed and accurate radiological findings. By adopting this approach, the system ensures that the generated reports align closely with those produced by radiologists. Furthermore, the use of prompt engineering enables customization of the reports to meet specific clinical needs. Ultimately, this project aims to improve report accuracy while simultaneously reducing the time healthcare professionals spend on administrative tasks, thereby streamlining the radiological reporting process.

II. Method

This project proposes a novel approach to generating radiology reports by framing the task as a retrieval problem. The process begins with a medical corpus of Chest X-ray reports, denoted as $C = \{c_1, c_2, \ldots, c_n\}$. When presented with an input image i, the system retrieves the top K relevant reports from this corpus. These retrieved reports and user and system prompts are then fed into Mistral AI's Retrieval-Augmented Generation (RAG) framework. This stage effectively serves as data augmentation, combining the domain-specific knowledge acquired from the retrieval task with the general expertise of the generative model. The result is a final report that benefits from both the targeted, relevant information in the retrieved reports and the broader language understanding capabilities of

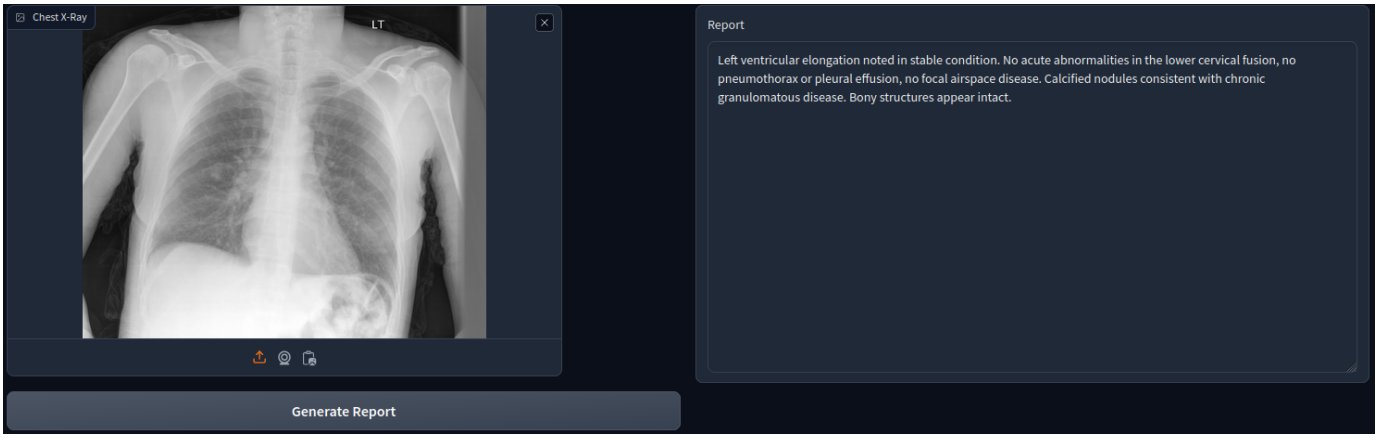

Figure 1: Web-based application created with Gradio, demonstrating report generation using MistralAI-XRay on a sample Chest X-ray image.

the AI model. This approach aims to produce more accurate, context-aware radiology reports by leveraging the strengths of both retrieval-based and generative AI methods.

A. Dataset

The open-access Indiana University Chest X-ray dataset, also known as the Open-I dataset, is used to train and test the model. This comprehensive collection includes 7,466 chest X-ray images and 3,851 corresponding radiology reports. Some reports correspond to both frontal and lateral X-ray images, while others correspond only to frontal X-rays. The dataset is sourced from two large hospital systems within the Indiana Network for Patient Care database. Each report is uniquely indexed with uid and typically includes sections such as indication, comparison, findings, and impression. The "Findings" section details the observations made from the X-ray images, while the "Impressions" section provides a concise summary of the diagnosis.

Upon analyzing the dataset, it is found that some reports lack the findings or impressions sections. During preprocessing, these issues are addressed by moving the impression data to the findings section if the findings are missing. Reports missing both sections are removed from the dataset. Additionally, the reports are processed to remove any unidentified words. After preprocessing, 7,426 images and 3,826 unique reports are retained. The same report is assigned to both the frontal and lateral images, resulting in 7,426 image-report pairs where reports comprise text from the findings section. These pairs are then divided into training, validation, and test sets in an 80:10:10 ratio.

B. Vision Language Model (VLM)

In this MistralAI-XRay project, ALBEF (Align Before Fuse) VLM model [6], is used as the backbone architecture for learning multi-modal embeddings. This sophisticated model enhances vision-language representation learning by aligning image and text representations before fusing them through cross-modal attention. ALBEF's architecture comprises three

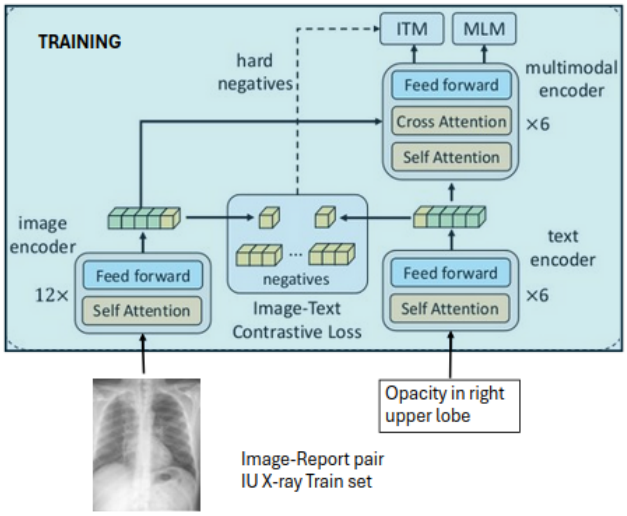

Figure 2: MistralAI-XRay Training Pipeline. Training set of IU X-ray is passed through ALBEF to learn joint embeddings of image-report pair.

essential components: a 12-layer Vision Transformer (ViT-B/16) image encoder with 85.8 million parameters, a text encoder based on the first 6 layers of BERT base with 123.7 million parameters, and a multimodal encoder using the last 6 layers of BERT base [10]. The multimodal encoder fuses outputs from the image and text encoders to generate image-text matching scores. ALBEF's design allows for direct computation of cosine similarity between input images and text, significantly improving the coherence and relevance of generated outputs. The model employs image-text contrastive learning (ITC) to align unimodal representations of image-text pairs before fusion, fostering more grounded vision-language representations. Additionally, ALBEF utilizes masked language modeling (MLM) and image-text matching (ITM) on the multimodal encoder to achieve superior joint representation of image-text pairs.

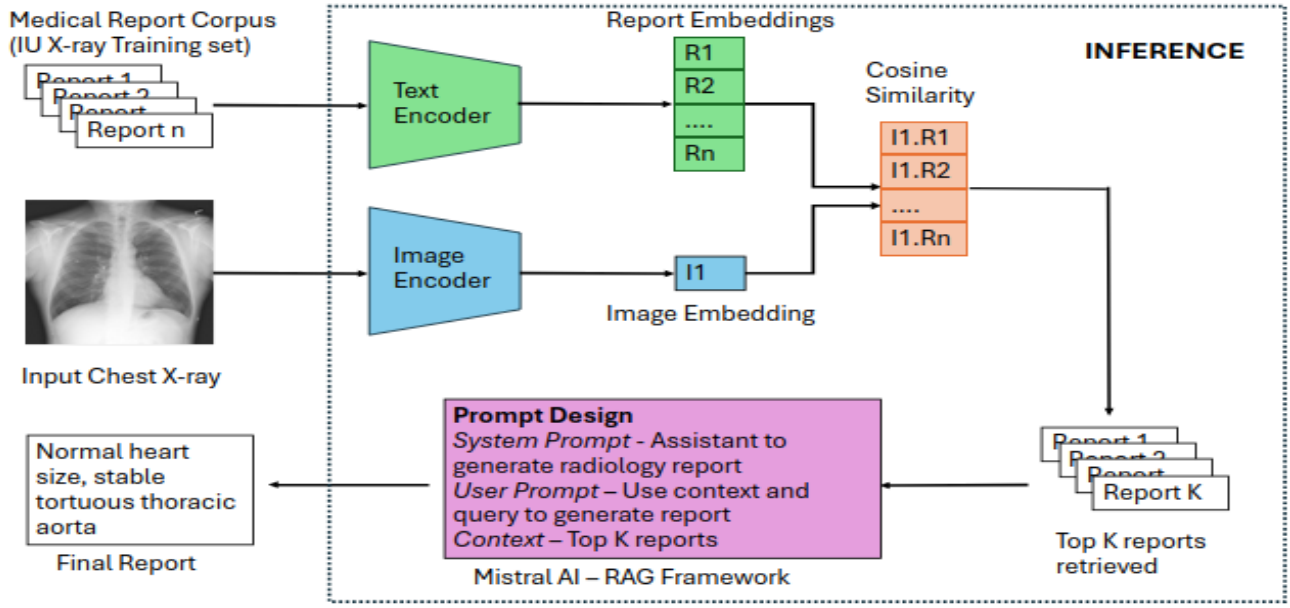

Figure 3: Web-based application created with Gradio, demonstrating report generation using MistralAI-XRay on a sample Chest X-ray image.

1) Training Phase- Learning Joint embeddings: Initially, the available pre-trained weights of ALBEF are uploaded, and then the model is trained on the IU X-ray dataset for 60 epochs. As part of the preprocessing, each input X-ray image is resized to 256x256 pixels and normalized. During training, the model is fed with image-text pairs in each iteration. The loss is computed using three key objectives: Image-Text Contrastive Learning (ITC), Masked Language Modeling (MLM), and Image-Text Matching (ITM). The model parameters are updated using the gradient descent method to minimize the loss, with AdamW as the optimizer. This training process allowed ALBEF to learn robust joint embeddings that accurately represent the relationships between chest X-ray images and their corresponding radiology reports.

2) Inference Phase- Reports retrieval: In the inference phase, generating radiology reports is approached as a retrieval task from a collection of reports denoted as $C = \{c_1, c_2, \ldots, c_n\}$. When given an input image of chest X-ray i, the ALBEF model is used to retrieve the most relevant report r from this collection. The process involves creating text embeddings for each report and an image embedding for the input X-ray using the unimodal components of ALBEF. The similarity score $s(c, i)$ is calculated as the dot product of the chest report embedding $t(c)$ and the image embedding $p(i)$. The report r is selected as the one that maximizes this similarity score, ensuring that the retrieved report is the most relevant and contextually appropriate for the given X-ray. Top K reports are retrieved from the inference phase, where hyperparameter $K = 3$ is set, which is then further processed.

C. Retrieval-Augmented Generation (RAG)

Once the top K relevant reports are retrieved, prompt engineering is utilized with Mistral AI's open-mistral-7b model, which is loaded and run using Ollama, incorporating the RAG framework to generate the final radiology report. This approach combines the strengths of retrieval-based methods with generative models to produce coherent and contextually accurate reports.

In this process, role-playing techniques are employed to guide the model effectively. A system prompt and a user prompt are designed to obtain the report using context and query. Table I shows the prompt engineering used to instruct the model to generate reports. The system prompt instructs the model to act as an assistant in generating a radiology report based on the provided context, where context is the top K relevant reports. This ensures the model understands its role and focuses on creating a relevant report. The retrieved documents are combined into a single context, which provides a comprehensive background for the report generation. The user prompt then incorporates this context along with a specific query related to the chest X-ray, asking the model to generate a detailed radiology report.

This methodology ensures that the model produces a coherent and detailed radiology report based on the context from the retrieved reports. The generated report is accurate, contextually relevant, and avoids any hallucinations with a well-crafted query, leveraging the retrieval-augmented generation technique. The overall report generation process is thus enhanced by combining the specific information retrieved from the corpus with the generative capabilities of Mistral AI, resulting in high-quality radiology reports tailored to the specific needs of the query and context provided.

Table I: Prompts used in RAG for generating radiology reports.

System Prompt	User Prompt	Query
You are an assistant designed to generate radiology reports.	Based on the following context and query, generate a detailed report: Context: {context} Query: {query}	Please generate a concise radiology report with only the impression section using information only from the retrieved context. ## Follow these guidelines to generate the report: 1) Do not provide the report in bullet points or numbers. 2) Do not write the word "impression," provide only the report. 3) Do not mention any comparison with prior or earlier reports. 4) Always give a single line report and don't introduce new line characters. 5) Do not provide any special characters. ##

D. Web-based App

A user-friendly web-based application has been developed using Gradio to facilitate the generation of radiology reports from chest X-ray images. The application allows users to upload a chest X-ray image, which is then processed using the MistralAI-XRay model to generate a detailed radiology report. Figure 1 shows the web app running and generating report. The interface is designed for simplicity and ease of use, making it accessible to both medical professionals and researchers. Upon uploading an image, the application leverages the underlying AI model to retrieve relevant reports from the training corpus, which are then used to generate a concise and accurate final report. The entire process is automated, providing users with a seamless experience from image upload to report generation.

III. Evaluation

The performance of the model is evaluated using the test set of the IU Chest X-ray dataset. The evaluation focuses on the final reports generated using Mistral AI. The model is evaluated using three key metrics: BLEUScore, BERTScore, and Semb, across top K report retrievals with K = 1, 2, 3.

BLEUScore (Bilingual Evaluation Understudy) [8] is a metric that evaluates the quality of machine-generated text by measuring n-gram overlap with reference reports. In radiology report generation, BLEU-2 is used, which assesses the overlap of unigrams and bigrams between the generated and reference reports. BERTScore [9] leverages BERT embeddings and assesses the semantic similarity between generated and reference reports. It compares their contextual embeddings to compute token-level similarity, focusing on semantic rather than direct token matches. Semb (Semantic Embedding) involves embedding the text into a high-dimensional space and then measuring the cosine similarity between the generated and reference reports. This metric helps in understanding the overall semantic alignment and coherence of the generated text with the reference, ensuring that the generated reports are not only lexically but also contextually relevant and accurate. Table II presents the evaluation metrics for MistralAI-XRay, with a BERT score of 0.34 and a Semb score of 0.42, indicating that the model performs well in accurately generating radiology reports.

Table II: Evaluation metrics for different values of K.

Metrics	K=1	K=2	K=3
BLEU	0.09	0.09	0.09
BERT	0.33	0.33	0.34
Semb	0.42	0.42	0.42

A. Ablation Study

An ablation study is conducted to understand the impact of varying the number of top K retrieval reports on the performance of the model for the report generation task. This analysis helps in determining the optimal number of retrieved reports that contribute to the most accurate and contextually relevant final radiology report.

1) Findings: The evaluation metrics II across different values of K (K=1, K=2, K=3) show minimal variation, with BLEU, BERT, and Semb scores remaining largely consistent. This stability suggests that the model performs reliably regardless of the number of retrieved reports used in the generation process. Specifically, the BERT score slightly improves from 0.33 to 0.34 as K increases, while the Semb score remains constant at 0.42, and the BLEU score remains at 0.09 across all K values.

These results indicate that the model effectively utilizes the context provided by a single report (K=1) and that adding additional reports (K=2 or K=3) does not significantly alter the performance. The consistent metrics across different K values highlight the robustness of the model's retrieval-augmented generation approach, ensuring that it maintains accuracy and coherence even when varying the amount of contextual information.

IV. Result

The analysis of reports generated by CXR-ReDonE and MistralAI-XRay shows that the retrieval-augmented generation (RAG) approach using Mistral AI produces significantly more concise and less noisy impressions compared to the pure retrieval-based strategy. As shown in Table III, which compares reports from Ground Truth, CXR-ReDonE, and MistralAI-XRay, the same-colored text indicates matches in each report. Mistral AI successfully retains all relevant clinical entities, ensuring that the reports are both accurate and concise.

Table III: Comparison of radiology report between Ground Truth, CXR-ReDonE, and MistralAI-XRay.

Ground Truth	**CXR-ReDonE**	**MistralAI-XRay**
The cardiomediastinal silhouette is within normal limits for size and contour. The lungs are normally inflated without evidence of focal airspace disease, pleural effusion, or pneumothorax. No acute bone abnormality.	Stable cardiomediastinal silhouette with normal heart size, mediastinal calcifications suggest a previous granulomatous process. Apical irregularities also present on the previous exam suggestive of scarring. No focal alveolar consolidation, no definite pleural effusion seen. No typical findings of pulmonary edema. No pneumothorax. The heart size and cardiomediastinal silhouette are normal. The aorta is tortuous and atherosclerotic. The lungs are hyperexpanded with flattening of hemidiaphragms and increased retrosternal airspace. There is no focal airspace opacity, pleural effusion, or pneumothorax. There are degenerative changes in the thoracic spine. Normal heart size. Stable tortuous thoracic aorta. Prior granulomatous disease. Healed rib fractures appear stable. Focal opacity is noted in the left midlung overlying the 9th posterior rib which represents healing rib callus. No pneumothorax or pleural effusion.	Cardiomediastinal silhouette and heart size are stable, normal. No evidence of pneumothorax or pleural effusion, no focal airspace disease. Calcified mediastinal nodules suggest chronic granulomatous process, apical irregularities consistent with scarring. Bony structures appear intact.
Stable appearance of lower cervical fusion Heart size normal. No pneumothorax or pleural effusion. No focal airspace disease. Calcified nodules consistent with chronic granulomatous disease. Bony structures appear intact. DISH of the thoracic spine.	Normal heart size. Stable tortuous thoracic aorta. Prior granulomatous disease. Healed rib fractures appear stable. Focal opacity is noted in the left midlung overlying the 9th posterior rib which represents healing rib callus. No pneumothorax or pleural effusion. Stable including elongation of the left ventricle and tortuous thoracic aorta. Subcarinal calcified lymph lung volumes. No focal consolidation, pneumothorax or large pleural effusion. Negative for acute bone abnormality. Low lung volumes bilaterally with central bronchovascular crowding without focal consolidation, pleural effusion, or pneumothoraces. Cardiomediastinal silhouette is within normal limits. Degenerative changes of the thoracic spine.	Stable thoracic aorta, healed left midlung rib callus over 9th posterior rib, normal heart size, no pneumothorax or pleural effusion, no focal airspace disease, calcified nodules consistent with chronic granulomatous disease, intact bony structures.

In contrast, the CXR-ReDonE method, which relies solely on retrieval, often introduces hallucinated text (highlighted in red) and frequently repeats phrases, making some information redundant. The RAG approach effectively filters out these insignificant details, which often clutter reports generated by pure retrieval methods, thereby improving overall readability and clinical utility. On the other hand, the baseline pure retrieval method tends to include extraneous details and exhibits higher noise levels, resulting in more verbose reports that reduce their clinical usefulness.

V. Conclusion

The combination of the Vision-and-Language Model (ALBEF) and the Retrieval-Augmented Generation (RAG) framework with Mistral AI has proven effective for automated radiology report generation. The ALBEF model, trained on the IU Chest X-ray (Open-I) dataset, creates robust joint embeddings of chest X-ray images and reports. During inference, relevant reports are retrieved and used as context for generating final reports through Mistral AI, utilizing advanced prompt engineering. These generated reports are more concise and less noisy than those from pure retrieval-based strategies, while still retaining all essential clinical information. An ablation study justified that the method is robust and does not get affected much by number of reports retrieved. Evaluation metrics confirmed the performance of MistralAI-XRay. This project confirms the potential of hybrid AI frameworks like ALBEF and RAG with Mistral AI to enhance the quality and efficiency of automated radiology report generation, supporting radiologists and improving patient care. Future work will focus on refining models, exploring additional datasets, and expanding the system's capabilities to more clinical scenarios.

References

[1] Michael P. Hartung, Ian C. Bickle, Frank Gaillard, and Jeffrey P. Kanne. How to create a great radiology report. *RadioGraphics*, 40(6):1658–1670, 2020.

[2] Ramesh, V., Chi, N., & Rajpurkar, P. (2022). Improving Radiology Report Generation Systems by Removing Hallucinated References to Non-existent Priors. *ML4H@NeurIPS*.

[3] Endo, M., Krishnan, R., Krishna, V., Ng, A.Y., & Rajpurkar, P. (2021). Retrieval-Based Chest X-Ray Report Generation Using a Pre-trained Contrastive Language-Image Model. *ML4H@NeurIPS*.

[4] Yu, F., Endo, M., Krishnan, R., Pan, I., Tsai, A., Reis, E.P., Fonseca, E.K., Lee, H.H., Abad, Z.H., Ng, A.Y., Langlotz, C., Venugopal, V.K., & Rajpurkar, P. (2022). Evaluating progress in automatic chest X-ray radiology report generation. *Patterns*, 4.

[5] Ranjit, M.P., Ganapathy, G., Manuel, R.F., & Ganu, T. (2023). Retrieval Augmented Chest X-Ray Report Generation using OpenAI GPT models. *ArXiv, abs/2305.03660*.

[6] Li, J., Selvaraju, R.R., Gotmare, A., Joty, S.R., Xiong, C., & Hoi, S.C. (2021). Align before Fuse: Vision and Language Representation Learning with Momentum Distillation. *Neural Information Processing Systems*.

[7] Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., Casas, D.D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L.R., Lachaux, M., Stock, P., Scao, T.L., Lavril, T., Wang, T., Lacroix, T., & Sayed, W.E. (2023). Mistral 7B. *ArXiv, abs/2310.06825*.

[8] Papineni, K., Roukos, S., Ward, T., & Zhu, W. (2002). Bleu: a Method for Automatic Evaluation of Machine Translation. *Annual Meeting of the Association for Computational Linguistics*.

[9] Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., & Artzi, Y. (2019). BERTScore: Evaluating Text Generation with BERT. *ArXiv, abs/1904.09675*.

[10] Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *North American Chapter of the Association for Computational Linguistics*.