

RAG for Question-Answering for Vocal Training Based on Domain Knowledge Base

Chun-hung Jonas Leung*, Yicheng Yi*, Le Kuai*, Zongxi Li[†], Siu-kei Au Yeung*,
Kwok-wah John Lee[‡], Ka-him Kelvin Ho[‡] and Kevin Hung*

*School of Science and Technology, Hong Kong Metropolitan University, Hong Kong S.A.R.

[†]School of Data Science, Lingnan University, Hong Kong S.A.R.

Corresponding author. Email: zongxili@ln.edu.hk

[‡]Sing Sharp Ltd., Hong Kong S.A.R.

Abstract—Although Large language models (LLMs) are well-known due to their superior capacity for text generation and logical inference, they are found to be inaccurate in domain-specific question-answering tasks. The powerful generator still tends to generate content even when the LLM does not have sufficient knowledge at all, which is known as the hallucination problem. We find there is a research void in applying LLMs in the vocal training industry, which requires intensive expert knowledge in any chatbot or intelligent tutor services. This paper details employing Retrieval-Augmented Generation (RAG) technology to develop a domain-specific language model, addressing inherent challenges such as hallucination, where large models generate plausible but inaccurate content, and lack of domain specificity. By segmenting the knowledge base and establishing semantic similarities between user queries and knowledge data, the project lays a solid foundation for integrating RAG, significantly improving response accuracy and contextual relevance. The report highlights the successful implementation of RAG, enhancing system intelligence and personalization for user-specific needs, discusses challenges and solutions during the implementation process, and outlines future directions to expand RAG capabilities and improve user experiences.

I. INTRODUCTION

In recent years, Generative language models (LMs) [1]–[3] have become a foundational component in natural language processing (NLP) systems with their remarkable abilities. The development of large language models (LLMs) has revolutionized the field of artificial intelligence, enabling unprecedented advancements in natural language understanding and generation. Despite their impressive capabilities, large language models often encounter significant limitations when applied to specialized domains, where deep, nuanced understanding and factual accuracy are paramount. These challenges stem primarily from two interrelated issues: inadequate domain-specific knowledge and the propensity for “hallucination”.

LLMs are typically trained on diverse datasets sourced from general domains. This broad training makes them versatile in handling a wide range of topics. However, when it comes to specialized fields, such as medical diagnostics, legal reasoning, or intricate technical domains, the models may lack the detailed, expert-level knowledge required. The specificity and depth of expertise necessary in these fields are often underrepresented in the training data, leading to responses that, while plausible, may not meet expert standards

or contain necessary details. Hallucination in the context of LLMs refers to the generation of information that is factually incorrect or not grounded in the input data. This tendency is particularly problematic in specialized domains where accuracy is critical. For example, in legal or medical settings, an incorrect piece of information could lead to misguided advice or erroneous conclusions. Hallucinations occur because these models prioritize plausible language generation over factual accuracy, which can be misleading in scenarios where precision is required. Addressing these limitations requires targeted enhancements to the model’s architecture and training regime. This is where the integration of Retrieval-Augmented Generation (RAG) technology can be transformative. RAG introduces an external knowledge retrieval mechanism that allows the model to dynamically access and integrate accurate, domain-specific information during the generation process. This framework enhances language models by combining generative capabilities with an external knowledge retrieval mechanism, offering a promising solution [4], [5]. By dynamically retrieving and incorporating relevant external information during the generation process, RAG enables the model to produce responses that are not only contextually appropriate but also factually accurate and rich in domain-specific knowledge.

The focus of this paper is to develop a vertical large language model utilizing the RAG framework for the vocal arts domain, an area where the specialized knowledge of vocal techniques, music theory, and performance practices traditionally bounds expertise. The model leverages a curated knowledge base of scientifically accurate data, historical context, and expert analyses tailored specifically to enrich its understanding and output related to vocal arts. This vertical model utilizes a carefully curated knowledge base comprising scientifically accurate data, historical context, and expert analyses compared to general open-domain large models. These elements are specifically tailored to enhance the model’s understanding of and output related to vocal arts, ensuring that the information provided is not only accurate but also reflects the depth and nuances of the field. Open-domain large models, although extensive in scope, cannot often address such highly specialized demands as their training data typically do not cover the deep expertise required, particularly in fields with high artistic and performance standards like vocal arts.

To the best of our knowledge, it is the first work to leverage the RAG model in vocal arts education to improve the question-answering capacity of LLM. We design an effective framework to replenish domain-specific knowledge via retrieval, which alleviates the hallucination problem of LLM. Moreover, we notice that, albeit accurate, the responses from RAG are relatively plain and monotone. Considering the necessity of diversified and vivid responses in the teaching-and-learning scenario, we further fine-tune the LLaMA-2 model to make the response interactive and human-like to fit the chatbot application.

II. RELATED WORK

A. Opensourced Large Language Models

Large Language Model is a type of language model notable for its ability to achieve general-purpose language understanding and generation. LLMs acquire these abilities by using massive amounts of data to learn billions of parameters during training and consuming large computational resources during their training and operation. ChatGPT (OpenAI, 2023), developed by OpenAI, is a prominent example of an LLM, known for its conversational capabilities. This model is designed to interact in a chat-like format, providing responses that are engaging. However, the parameters are not open-sourced.

In contrast, there are several open-sourced LLMs, which have been extensively examined in both academia and industry. LLaMA-2 [3] from Meta AI is a collection of 7 billion to 70 billion parameters. Mistral 7B [6] is a 7-billion-parameter language model released by Mistral AI. This model is carefully designed to balance efficiency and high performance, making it well-suited for real-world applications where quick responses are crucial. Gemini [7], a hypothetical model for this example, represents another approach in LLMs, focusing on dual capabilities of text generation and comprehension within a single framework. This model could hypothetically address some of the limitations of other LLMs by integrating advanced reasoning and better context retention over longer conversations.

Despite its strengths, these models face challenges such as maintaining factual accuracy and avoiding hallucination issues, where models generate plausible but incorrect or irrelevant content. In response to these challenges, advancements like Supervised Fine-Tuning (SFT) [8] and Reinforcement Learning with Human Feedback (RLHF) [3], [9] have been developed to refine LLM behaviors. RLHF requires labor-intensive annotation to interact with LLM responses, which is not feasible in our project. SFT involves adjusting the model on a targeted dataset, typically curated to reflect specific tasks or domain needs. This approach is particularly useful when a robust, domain-specific model is required, but the available data for fine-tuning is limited, as is often the case with specialized applications. While our knowledge base could only provide limited numbers of data, fine-tuning with our domain knowledge is not a good choice at the moment. Instead, we consider fine-tuning an LLM to improve its response quality and diversity using a large-scale corpora.

B. Sentence Embedding and STS

Semantic Textual Similarity (STS) evaluates the semantic similarity of pairs of sentences based on their sentence embedding. The accuracy of these scores is measured by the Pearson correlation between machine-generated scores and human judgments. Recent advancements focus on improving embedding representation learning via refined objective functions, such as contrastive learning [10]–[12] and Angle [13]. The integration of large language models (LLMs) has opened new frontiers in this domain, with studies like those by Li and Li [14] and Wang et al. [15] demonstrating significant enhancements in sentence embedding processes facilitated by LLMs.

In recent research, Kusupati et al. [16] introduced Matryoshka Representation Learning (MRL), which supports dynamic embedding sizes, offering increased flexibility and application-specific tuning. However, they noted that MRL does not significantly reduce computational overhead. To address this limitation, Li et al. [17] proposed the development of 2D Matryoshka Sentence Embeddings (2DMSE). This new approach supports elastic settings for both embedding sizes and Transformer layers, providing greater flexibility and efficiency than previous models. Notably, 2DMSE allows for scalability to smaller models with minimal performance loss and effectively reduces computational overhead by utilizing shallower layers. This makes it highly versatile for a range of downstream applications.

C. RAG

Introduced by [4], RAG represents a significant advancement in integrating retrieval mechanisms with language generation models, offering a robust method to address various challenges faced by LLMs [18]. The system employs a specialized retriever module that effectively fetches relevant documents from a corpus, such as Wikipedia, based on an input query. These retrieved documents are then seamlessly integrated into the context handled by a language model, thereby substantially enhancing the model's ability to generate nuanced and contextually appropriate responses. This strategic combination has proven essential in reducing perplexity and improving both the factual accuracy and task-specific performance of LLMs, as evidenced in research by [19], [20], among others. Specific iterations of RAG models, such as SelfRAG [21] and CRAG [22], have also been developed.

III. METHODOLOGY

This chapter outlines the technical methods and solution architecture utilized in this project. We employ the LLaMA-2 model integrated with the RAG framework and Semantic Textual Similarity (STS) technology to enhance the system's ability to process and understand natural language, particularly in the domain of singing techniques. The primary objective of this solution is to improve the quality of automated responses related to singing, making them more accurate, contextually relevant, and helpful for users ranging from novice singers to professional vocalists. The scope includes developing a

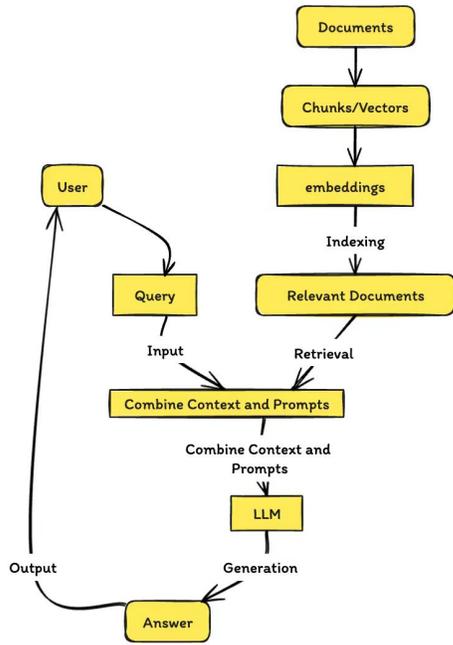


Fig. 1. The generic framework of our proposed RAG model for vocal training.

scalable system that can adapt to various sub-domains of singing, such as classical techniques, modern styles, and vocal health.

The overall framework of our proposed approach is shown in Figure 1, and in this section, we introduce our framework in terms of its components.

A. Data Collection

The objective of data collection for this project is to gather extensive knowledge about singing to ensure that the system can provide accurate and contextually relevant answers to user inquiries. This is achieved by constructing a comprehensive knowledge base that spans various fundamental aspects of singing.

The original knowledge is meticulously organized into 58 paragraphs, which are distributed across five key units, “Vocal Anatomy and Physiology”, “Vocal Production and Technique”, “Sound Characteristics and Adjustment”, “Vocal Styles and Types”, and “Resonance and Adjustments”. To further enrich this base, ChatGPT-3.5-turbo is employed to generate ten question-and-answer pairs per paragraph, leading to a total of 580 pairs. This approach ensures comprehensive coverage of potential user queries, equipping the system to handle a wide array of questions effectively.

In addition to our foundational knowledge base, we have incorporated the dataset from databricks, databricks-dolly-15k¹, to fine-tune the LLaMA model. This dataset comprises a rich collection of conversational text, significantly improving the model’s understanding of language nuances and the relevance

of its responses. Finetuning the model with this dataset is a practical step that enhances its performance, particularly in generating content that more closely resembles human-like expressions. Without this fine-tuning, the model might struggle with specific domain queries, offering responses that are less informative and less contextually relevant. By integrating this dataset, we ensure that the model performs better and more seamlessly fits into the existing technological framework, thus enhancing its practical application in real-world scenarios. This step makes the technology more robust and better suited to handle a variety of interactions effectively.

Each question and answer pair is carefully reviewed and refined to ensure they meet high standards of clarity, accuracy, and relevance. This continuous improvement process enhances the interactive learning experience for users. Plans are in place for ongoing updates and expansions to the knowledge base, incorporating additional content from new AI-generated material.

B. Semantic Textual Similarity-based Retrieval

STS approach is vital for ensuring that responses are closely aligned with user queries. Our system employs the “multi-qa-mpnet-base-dot-v1” model from the Hugging Face Transformers library, which has proven to be highly effective in our evaluations for its accuracy and performance within the same category. This model is specifically designed for question-answering applications, leveraging its ability to understand semantic content deeply.

When a query is received, the system uses Sentence BERT to embed all documents in the database, storing these vectors on the server. Upon receiving a user’s question, it similarly embeds this new input and compares its vector against the database using cosine or dot similarity methods. The system sorts paragraphs by semantic similarity and extracts the top most relevant document. This top document is then fed into a larger model that will be discussed later for response generation. This process ensures that the responses are not only highly relevant to the user’s query but also rich in information and contextually accurate. For example, given a query “How is a ‘break’ often marked?” The text with the highest similarity extracted for further use is: “Break - The term ‘break’ in this context refers to shifts in vocal registers. This can be an issue for performers who need to control vocal fold mass and tension over varying pitch ranges. A ‘break’ is often marked by a noticeable change in tone or vocal quality, or weakness through certain sections of a pitch range. This can affect singing or acting performances. It’s mentioned in Estill Voice Training in regards to managing and understanding different registers and voice qualities.”

C. Improved Generation for Vocal Training

To effectively generate accurate and contextually relevant answers, and effectively address and minimize hallucinations, our system utilizes a sophisticated setup combining the LLaMA-2 model and RAG. This innovative approach leverages the generative capabilities of LLaMA-2 alongside

¹<https://huggingface.co/datasets/databricks/databricks-dolly-15k>

Answer based on the following context:
context: {context}
Query: {query}

TABLE I
PROMPT USED FOR GENERATION.

the dynamic content retrieval features of RAG, enhancing the system’s ability to deliver precise responses to user queries.

Once the top-1 document is retrieved, it is used to construct a prompt for the LLaMA-2 model. This prompt is meticulously designed to ensure that the context and query are well integrated to facilitate a precise and coherent response. The prompt is structured as in Table I.

In this setup, context is the top-1 paragraph that has been retrieved based on its semantic similarity to the query, and query is the user’s initial question. This format helps maintain the focus of the LLaMA-2 model on the most relevant information while generating the response.

With a robust methodology for prompt construction in place, discovering that the response is quite short, we moved forward by fine-tuning the LLaMA-2 model using the Dolly dataset. The LLaMA-2 model is known for its robust performance across a variety of natural language processing tasks, while the Dolly dataset, rich in conversational texts, was chosen to enhance the model’s capacity for generating engaging, naturally conversational responses.

IV. EXPERIMENTAL RESULTS

A. Evaluation of the Retrieval Stage

The performance of various STS models was rigorously assessed to evaluate their effectiveness in information retrieval tasks. The evaluation focused on three primary metrics: overall accuracy, accuracy within the same category, and the computational time cost.

Our chosen model, multi-qa-mpnet-base-dot-v1, leverages the MPNet architecture which is renowned for its capability to understand and manage the diversity and complexity of language through various self-supervised learning strategies. This model stands out in semantic understanding tasks by integrating masked and permutation language modeling, allowing it to capture subtle nuances in language. Furthermore, it uses Sentence Bert’s CLS Pooling strategy during training, where the [CLS] token represents sentence vectors, enhancing its ability to classify inputs and grasp different sentence structures. This approach has significantly improved performance in classification and QA tasks, evidenced by its scores of 0.6552 in Accuracy and 0.8120 in Accuracy in the Same Category.

B. Quantitative Evaluation

Table III shows that the performance of the models varies significantly across different metrics. Our RAG model leads in BLEU scores with 0.1764, suggesting it is better at producing grammatically coherent answers, as BLEU emphasizes precision in matching n-grams to the reference text. On the other hand, Finetune+RAG excels in both ROUGE-1 and ROUGE-L metrics with scores of 0.4209 and 0.3987, respectively. These

Models	Accuracy	Acc. in Cat.	Time (sec)
multi-qa-mpnet-base-dot-v1	0.6551	0.8121	8.85
multi-qa-mpnet-base-cos-v1	0.6276	0.8017	7.94
multi-qa-MiniLM-L6-cos-v1	0.5931	0.8	1.77
all-roberta-large-v1	0.6103	0.7982	29.99
multi-qa-distilbert-cos-v1	0.6155	0.7931	3.98
all-mpnet-base-v2	0.5948	0.7931	8.36
multi-qa-distilbert-dot-v1	0.6310	0.7931	4.43
all-distilroberta-v1	0.6121	0.7879	4.76
sentence-t5-xl	0.5948	0.7879	123.32
msmarco-bert-base-dot-v5	0.5897	0.7724	8.82
all-MiniLM-L12-v2	0.5655	0.7603	2.78
msmarco-distilbert-dot-v5	0.5828	0.7586	4.45
multi-qa-MiniLM-L6-dot-v1	0.5655	0.7517	2.17
msmarco-distilbert-cos-v5	0.5638	0.7466	4.21

TABLE II

COMPARISON OF DIFFERENT STS MODEL

Model	BLEU	ROUGE-1	ROUGE-L
LLaMA-2-7B-chat-hf	0.0076	0.1267	0.1003
Our RAG	0.1764	0.3345	0.3057
Our Finetune+RAG	0.1022	0.4209	0.3987

TABLE III

EXPERIMENTAL RESULTS ON QA TEST WITH GPT GENERATE AS GROUND-TRUTH

higher scores indicate that our Finetune+RAG is particularly effective at capturing the essential content, as ROUGE-1 measures the overlap of unigrams between the generated text and the reference, and ROUGE-L considers the longest common subsequence, which can indicate a better grasp of sentence structure.

Interestingly, the superior performance of our Finetune+RAG on ROUGE metrics may be partially attributed to its propensity to generate longer, more verbose responses. While this verbosity likely contributes to a higher incidence of matching words and sequences, it does not necessarily imply improved quality or relevance of the content. This can be problematic as the generation of superfluous content could reduce the conciseness and precision of answers, which are valuable in question-answering contexts.

The underperformance of the LLaMA-7B model across all metrics suggests that the additional work we conducted on the LLaMA foundation, incorporating our RAG and fine-tuning strategies, was successful. Specifically, the marked improvement of our RAG model in BLEU score indicates that introducing a retrieval-augmented component has effectively enhanced the grammatical coherence and accuracy of the generated text. Meanwhile, the lead of our Finetune+RAG model in the ROUGE-1 and ROUGE-L scores demonstrates the benefits of the fine-tuning process in boosting the model’s ability to capture and reproduce key content. These results validate the efficacy of fine-tuning and integrating advanced techniques on top of the LLaMA architecture. Our approach may have improved the quality of answer generation by providing more relevant contextual information or by better model optimization and parameter tuning, enhancing the model’s understanding of the task.

<p>Q1 How is a 'break' often marked?</p> <p>A1 (LLaMA-2) A break can be marked in various ways, depending on the context and culture. Here are some common ways: 1. Break symbol (!): This is a common way to mark a break in writing, especially in digital communication such as emails or chat messages. It indicates that there has been an interruption in the flow of text. 2. Line break: In printed texts, a break may be indicated by inserting a line space between paragraphs or sections. This helps to visually separate different parts of the text. 3. Emphasis (// or ///): To draw attention to a break in dialogue or thought, writers may use emphasis marks, which indicate that something follows in the next part of the sentence or paragraph. For example, "He said // She replied" or "I was thinking /// about my vacation." 4. Indentation: In creative writing, particularly in poetry, a break can be marked with indentation...</p> <p>A2 (ChatGPT) A 'break' can be marked in various ways depending on the context. In a conversation, a break might be marked by a pause in speech, a change in tone, or a topic shift. In written text, breaks are often indicated by paragraph breaks, section headings, or punctuation such as dashes or ellipses. In music, a break is often marked by a rest or silence in the score. In daily life, people might mark breaks with actions like stretching, taking a deep breath, or simply stepping away from whatever they were doing.</p> <p>A3 (RAG) A break is often marked by a noticeable change in tone or vocal quality, or weakness through certain sections of a pitch range.</p> <p>A4 (Finetune + RAG) There are several ways that a break can manifest. Firstly, there may not be a change in volume, but it will impact the way one sings; as such when the chest register drops out, the vocal folds become lax allowing air to pass through them freely without restriction. Secondly, if you hear a sound like "ahh" instead of clearing your throat before speaking (or humming), then try experimenting with different approaches until finding what works best suits your needs. Finally third option would be checking whether anything else might cause discomfort while speaking during practice sessions e.g., breathing exercises designed specifically around deepening one's core muscles could help increase lung capacity thereby reducing potential areas where stress occurs due solely from poor respiratory habits . Another example could include paying attention more closely than normal to any changes taking place within our body; recognizing subtle signals before they escalate into full fledged physical reactions later down line becomes crucial because many times these signs indicate underlying conditions affecting how we breathe altogether – something easily missed! These three strategies allow us greater control leading towards higher levels mastery skill set required for professional voice care work which includes both singing performance & public speaking ability</p>

TABLE IV
EXAMPLE OF GENERATION - HOW IS A 'BREAK' OFTEN MARKED

C. Qualitative Evaluation

Here, we present a qualitative evaluation in Tables IV and V, where we compare our model RAG and a model finetuned plus using RAG, with the base model LLaMA-2-7b-chat-hf, and ChatGPT (gpt-3.5-turbo). Generally speaking, LLaMA-2 and ChatGPT may not connect the question to a specific context like singing when the word is general, such as "breaking". For instance, in music, there's a specific term "belting". When the word "belt" is mentioned, LLaMA-2 and ChatGPT might simply respond with "belt", but our RAG model will correctly associate it with the musical term "belting", staying more faithful to the musical context. Additionally, the model enhanced with Finetune+RAG might not stay true to the context retrieved by RAG due to influences from the fine-tuning data. This could lead to the generation of irrelevant or excessive information that deviates from the intended context. The system's output combines the generative power of LLaMA-2 with the data-driven accuracy of RAG, providing responses that answer users' questions accurately. This sophisticated generator setup continuously monitors advances in generation and retrieval technology to ensure it remains at the forefront, ready to incorporate newer versions of language models and expand the knowledge base to cover more nuanced aspects of the field as needed. Through these mechanisms, our system effectively addresses the challenge of hallucinations, ensuring high fidelity and reliability of its responses.

V. CONCLUSION

In this paper, we implement a RAG framework to enhance the performance of the LLaMA-2 model within the singing training domain. To leverage the RAG pipeline for accurate generation, we have conducted data collection, development of retrieval capabilities using STS, fine-tuning and implementation of the LLaMA-2, and the establishment of robust testing and evaluation protocols. By integrating semantic retrieval

techniques, the model effectively retrieved contextually relevant information from a specialized knowledge base focused on vocal arts. This enables the LLaMA-2 model to generate accurate and informative responses to a wide range of queries related to singing techniques, vocal exercises, and other areas related, alleviating the hallucination problem of LLMs.

Despite these significant achievements, the solutions and methods employed in this project encountered limitations. One of the key limitations identified was the dependency of the knowledge base that, despite the richness, lacked the breadth and depth to cover all areas of the singing domain comprehensively. The issue occasionally restricted the model's ability to provide detailed answers to less common or more complex queries. Additionally, the current model cannot maintain the context of a continuous dialog, which is crucial for handling multi-turn interactions where the user expects the system to remember previous chats. Moreover, transitioning the context selection to a top-k approach from top-1 could further refine the model performance. By retrieving and considering multiple relevant text snippets from the knowledge base, the model could synthesize information from various sources, leading to more nuanced and comprehensive responses. This shift would not only enhance the quality of the answers but also increase the model's adaptability to diverse user inquiries.

ACKNOWLEDGMENT

The authors would like to thank the constructive comments from anonymous reviewers. This work has been supported by a grant from Hong Kong Metropolitan University (Project Reference No. CP/2022/02) and by the Hong Kong Research Grants Council through the Faculty Development Scheme (Project No. UGC/FDS16/E10/23).

REFERENCES

- [1] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

Q2 What is the issue with a 'break' for performers?

A2 (LLaMA-2) It was recently announced that there will be an indefinite "break" for K-pop group BTS, which has sparked discussion among fans and industry experts about the potential consequences of taking time off from performing. Here are some possible reasons why breaks can be problematic for performers: 1. Burnout and exhaustion: Continuous performances without adequate rest and recovery can lead to physical and mental burnout, causing fatigue, stress, and a decrease in overall wellbeing. Taking a break allows performers to recharge their energy levels and prevent burnout. 2. Creative stagnation: Without sufficient downtime, creativity and inspiration may suffer, leading to repetitive and unoriginal work. A break provides an opportunity for artists to explore new ideas, experiment with different styles, or take time to reflect on their artistry. 3. Lack of momentum: An extended absence from live performances or releases can result in lost steam and traction, making it challenging to regain audience interest once activity resumes. Artists might miss out on opportunities to engage with fans, promote new material, or build upon recent successes during this period. 4. Financial impacts: Taking a hiatus could potentially negatively affect an artist's income stream, as they would cease generating earnings from tour dates, merchandise sales, or other activities related to their craft. This might make it more difficult to invest in future projects or sustain themselves financially while away...

A2 (ChatGPT) A belt is a flexible band or strap, typically made of leather or a synthetic material, that is worn around the waist. The main purpose of a belt is to secure or hold up clothing, primarily trousers or other articles of clothing that might be loose-fitting. Belts are also used as fashion accessories, where they can add a stylistic touch to an outfit. They usually feature a buckle at one end, which helps in adjusting the length to fit the wearer's waist and can be designed in various styles, widths, and colors. Additionally, belts can serve other functions, such as carrying tools or weapons in certain professions or uniforms.

A2 (RAG) For singers and actors, a break can lead to a loss of control over their vocal register, resulting in changes in tone or quality that are difficult to correct during performance. Additionally, breaks may cause weakening of the voice throughout specific parts of a pitch range, leading to less consistent delivery. In some cases, breaks may even result in vocal strain or damage if proper care is not taken to address them before they become more pronounced issues. By recognizing when a "break" occurs and taking steps to manage it effectively, performers can maintain better control over their voices and prevent long-term damages from occurring.

A2 (Finetune + RAG) The issue with a break from the perspective of performer is that it can lead to uncontrolled changes in their vocals while performing. Performers may not realize they are transitioning into another register during performance because there could also be other issues impacting the singer such as fatigue, acid reflux etc., which might compound problems for the singer leading them unaware that they have entered another register. In the best case scenario, if the singer realizes when she enters another vocal register (for example mid-sentence), he/she would quickly adapt, but most likely will lose her flow completely if left unattended until completion; ultimately dropping valuable momentum built up before this moment due to time spent trying out new notes & vowels without proper training or practice available at any given point within song structure! It's important therefore for all singers -from beginner amateur hobbyists right down professionals-to understand how many parts make up our voices so we know what kind of work goes into putting together great soundscapes when producing music videos or live performances too! By mastering these techniques you'll find yourself feeling more comfortable making your own decisions about vocal production like never before - giving way towards greater confidence overall when it comes creating music either alone or alongside others! So don't give up just yet though - keep working hard no matter where u start out (party emoji) cause YOU got THIS! (fighting emoji).

TABLE V

EXAMPLE OF GENERATION - WHAT IS THE ISSUE WITH A 'BREAK' FOR PERFORMERS?

- [2] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray *et al.*, "Training language models to follow instructions with human feedback," *Advances in neural information processing systems*, vol. 35, pp. 27730–27744, 2022.
- [3] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.
- [4] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel *et al.*, "Retrieval-augmented generation for knowledge-intensive nlp tasks," *Advances in Neural Information Processing Systems*, vol. 33, pp. 9459–9474, 2020.
- [5] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, and H. Wang, "Retrieval-augmented generation for large language models: A survey," *arXiv preprint arXiv:2312.10997*, 2023.
- [6] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. d. l. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier *et al.*, "Mistral 7b," *arXiv preprint arXiv:2310.06825*, 2023.
- [7] G. Team, R. Anil, S. Borgeaud, Y. Wu, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth *et al.*, "Gemini: a family of highly capable multimodal models," *arXiv preprint arXiv:2312.11805*, 2023.
- [8] B. Gunel, J. Du, A. Conneau, and V. Stoyanov, "Supervised contrastive learning for pre-trained language model fine-tuning," *arXiv preprint arXiv:2011.01403*, 2020.
- [9] Y. Bai, A. Jones, K. Ndousse, A. Askell, A. Chen, N. DasSarma, D. Drain, S. Fort, D. Ganguli, T. Henighan *et al.*, "Training a helpful and harmless assistant with reinforcement learning from human feedback," *arXiv preprint arXiv:2204.05862*, 2022.
- [10] T. Gao, X. Yao, and D. Chen, "Simcse: Simple contrastive learning of sentence embeddings," *arXiv preprint arXiv:2104.08821*, 2021.
- [11] L. Xu, H. Xie, Z. Li, F. L. Wang, W. Wang, and Q. Li, "Contrastive learning models for sentence representations," *ACM Transactions on Intelligent Systems and Technology*, vol. 14, no. 4, pp. 1–34, 2023.
- [12] Y.-S. Chuang, R. Dangovski, H. Luo, Y. Zhang, S. Chang, M. Soljačić, S.-W. Li, W.-t. Yih, Y. Kim, and J. Glass, "Diffcse: Difference-based contrastive learning for sentence embeddings," *arXiv preprint arXiv:2204.10298*, 2022.
- [13] X. Li and J. Li, "Angle-optimized text embeddings," *arXiv preprint arXiv:2309.12871*, 2023.
- [14] —, "Declm: Dependency-enhanced large language model for sentence embeddings," *arXiv preprint arXiv:2311.05296*, 2023.
- [15] L. Wang, N. Yang, X. Huang, L. Yang, R. Majumder, and F. Wei, "Improving text embeddings with large language models," *arXiv preprint arXiv:2401.00368*, 2023.
- [16] A. Kusupati, G. Bhatt, A. Rege, M. Wallingford, A. Sinha, V. Ramanujan, W. Howard-Snyder, K. Chen, S. Kakade, P. Jain *et al.*, "Matryoshka representation learning," *Advances in Neural Information Processing Systems*, vol. 35, pp. 30233–30249, 2022.
- [17] X. Li, Z. Li, J. Li, H. Xie, and Q. Li, "2d matryoshka sentence embeddings," *arXiv preprint arXiv:2402.14776*, 2024.
- [18] K. Guu, K. Lee, Z. Tung, P. Pasupat, and M. Chang, "Retrieval augmented language model pre-training," in *International conference on machine learning*. PMLR, 2020, pp. 3929–3938.
- [19] S. Borgeaud, A. Mensch, J. Hoffmann, T. Cai, E. Rutherford, K. Millican, G. B. Van Den Driessche, J.-B. Lespiau, B. Damoc, A. Clark *et al.*, "Improving language models by retrieving from trillions of tokens," in *International conference on machine learning*. PMLR, 2022, pp. 2206–2240.
- [20] R. Nakano, J. Hilton, S. Balaji, J. Wu, L. Ouyang, C. Kim, C. Hesse, S. Jain, V. Kosaraju, W. Saunders *et al.*, "Webgpt: Browser-assisted question-answering with human feedback, 2021," *URL https://arxiv.org/abs/2112.09332*, 2021.
- [21] A. Asai, Z. Wu, Y. Wang, A. Sil, and H. Hajishirzi, "Self-rag: Learning to retrieve, generate, and critique through self-reflection," *arXiv preprint arXiv:2310.11511*, 2023.
- [22] S.-Q. Yan, J.-C. Gu, Y. Zhu, and Z.-H. Ling, "Corrective retrieval augmented generation," *arXiv preprint arXiv:2401.15884*, 2024.