VoxRAG: A Step Toward Transcription-Free RAG Systems in Spoken Question Answering

Zackary Rackauckas[1], Julia Hirschberg[1],
[1]Columbia University,
zcr2105@columbia.edu, julia@cs.columbia.edu

Abstract

We introduce VoxRAG, a modular speech-to-speech retrieval-augmented generation system that bypasses transcription to retrieve semantically relevant audio segments directly from spoken queries. VoxRAG employs silence-aware segmentation, speaker diarization, CLAP audio embeddings, and FAISS retrieval using L2-normalized cosine similarity. We construct a 50-query test set recorded as spoken input by a native English speaker. Retrieval quality was evaluated using *LLM-as-a-judge* annotations. For very relevant segments, cosine similarity achieved a Recall@10 of 0.34. For somewhat relevant segments, Recall@10 rose to 0.60 and nDCG@10 to 0.27, highlighting strong topical alignment. Answer quality was judged on a 0–2 scale across relevance, accuracy, completeness, and precision, with mean scores of 0.84, 0.58, 0.56, and 0.46 respectively. While precision and retrieval quality remain key limitations, VoxRAG shows that transcription-free speech-to-speech retrieval is feasible in RAG systems.

1 Introduction

Traditional question-answering (QA) retrieval-augmented generation (RAG) systems retrieve text documents from a vector database by performing semantic similarity search from a user's query. A large language model (LLM) then generates context-aware answers based on the retrieved content (Rackauckas, 2024). This architecture, however, can be extended to operate directly on spoken audio instead of text. Retrieving spoken audio documents without relying on intermediate transcriptions is an emerging area of RAG research (Min et al., 2025).

We present VoxRAG: a modular, open-source retrieval pipeline for RAG with full speech-to-speech retrieval. Unlike hybrid text and audio systems, VoxRAG keeps both the user query and retrievable documents in audio form up to the generation stage, using Contrastive Language-Audio Pretraining (CLAP) embeddings (Elizalde et al., 2022) to retrieve semantically relevant segments directly from podcast audio (see Appendix D for sample QA pairs).

Using podcasts as a retrieval target presents challenges such as informal language, overlapping speakers, non-speech audio (e.g., music, laughter), and generally poor automatic speech recognition (ASR) transcription output quality (Jones et al., 2021). VoxRAG mitigates these issues with silence-aware segmentation, speaker diarization, and CLAP embedding retrieval, avoiding early commitment to potentially faulty transcripts and enabling semantically grounded retrieval in the acoustic domain. We evaluate both retrieval and answer quality using RAGElo's *LLM-as-a-judge* methods, which have shown positive alignment with human judgments in QA evaluation, to assess how well retrieved audio supports answer generation (Rackauckas et al., 2024).

Related work has explored RAG systems for audio in both text and hybrid modalities. The TREC 2020–21 Podcasts Track saw systems using ASR text retrieval and summarization (Clifton et al., 2020), including fine-tuned BART (Lewis et al., 2019) and Whisper spoken term detection. Hybrid systems like Schwertzer's combination of COLA (Saeed et al., 2020) and RoBERTa (Liu et al., 2019) show promise in mixed-modal retrieval (Schwerter, 2022).

More recent models embed audio and text into shared or comparable vector spaces. SpeechDPR distills from ASR and dense passage retrieval (DPR) systems to embed spoken passages directly (Lin et al., 2024), while SEAL uses separate encoders for speech and text to enable cross-modal retrieval without transcription (Sun et al., 2025). Spectron processes spectrograms for QA entirely within an LLM framework (Nachmani et al., 2024), and SpeechRAG integrates speech retrieval with

Proceedings of the 1st Workshop on Multimodal Augmented Generation via Multimodal Retrieval (MAGMaR 2025), pages 40–46
August 1, 2025

an LLM for answering text queries from raw audio (Min et al., 2025). Meanwhile, DUAL demonstrates fully speech-native retrieval by embedding discrete speech units without paired text training (Lin et al., 2022). VoxRAG contributes to this emerging space by exploring a retrieval-first, speech-native architecture that maintains audio representations up to the point of answer generation, differing from span-prediction models like DUAL.

2 Method

Our construction of VoxRAG was motivated by two core research questions: 1) Can we retrieve semantically relevant documents directly from spoken language and without relying on text representations? 2) Can those documents support high-quality answer generation using an LLM?

We define "high-quality" segments as those that contain very relevant or somewhat relevant information to a user's query. We define "high-quality" answers along four axes: relevance, accuracy, completeness, and precision.

2.1 Podcast Indexing

Each podcast is processed through a modular indexing pipeline with speaker diarization, silence-aware segmentation, audio embedding, and optional transcription. Diarization is handled via NeMo's ClusteringDiarizer (Kuchaiev et al., 2019), while speech segmentation uses Silero VAD (Silero Team, 2024). Transcripts are generated using Faster-Whisper (Radford et al., 2022; SYSTRAN) and are only used for LLM input and display rather than retrieval.

All speech segments are embedded using CLAP (Elizalde et al., 2022), which maps audio to a joint audio-language embedding space (see Appendix C). This allows semantic-level retrieval even in the absence of exact word overlap, making it more robust for podcast audio that includes informal speech, background noise, or laughter. While traditional models like *wav2vec 2.0* (Baevski et al., 2020) focus on phonetic or acoustic information, CLAP learns to associate audio with language in a shared space. This lets us treat podcast segments like paragraphs of meaning rather than waveforms or phonemes (Elizalde et al., 2022) for direct audio retrieval.

Audio Loading and Preprocessing: Each podcast file is loaded, converted to mono, and resampled to 16 kHz.

Segmentation and Diarization: Diarization is used to detect speaker turns and assign segment-level speaker IDs. VAD identifies valid speech spans, which are then merged with speaker labels to define segments.

Embedding and Optional Transcription: Segments are embedded with CLAP and stored in memory. Transcripts are generated and aligned with segments for LLM prompting.

2.2 Retrieval

At query time, we take a spoken user query, process it through the same pre-processing and CLAP embedding pipeline, and compute cosine similarity between the query and all indexed segment embeddings using FAISS (Douze et al., 2025) (see Appendix C). The top ten segments are selected as candidates. We evaluate two retrieval configurations: (i) cosine similarity only and (ii) cosine followed by the *ms-marco-MiniLM-L6-v2* cross-encoder reranker. All other hyper-parameters are kept identical. Our primary analysis focuses on retrieval using cosine similarity, as shown in Table 1.

Query Processing: The user's spoken query is loaded, normalized, and embedded using CLAP.

Similarity Search: The top ten segments are retrieved by cosine similarity. Neighboring segments (before and after) are included for context.

2.3 Answer Generation

VoxRAG's modularity supports evaluation of chunking, embedding, and retrieval strategies. Once segments are retrieved, their transcripts are passed along with the transcribed query to *GPT-4o* to generate a natural language response.

Prompt Construction: Retrieved segment transcripts are labeled with the speaker and the segment number. The transcribed query and these segments are formatted as a prompt for the LLM.

Generation and Display: *GPT-4o* returns a natural language answer. The answer is shown in a Gradio interface alongside audio players for each segment.

3 Experiments

3.1 Dataset and Evaluation Queries

We selected twenty episodes from the Trash Taste podcast as our source corpus. These episodes feature three main speakers: Joey, Connor, and Garnt, with occasional guest speakers. For our main evaluation, we used a single representative episode with a run time of 2 hours and 3 minutes. This episode was segmented into 202 chunks using silence-aware merging and speaker diarization, ensuring that each segment remained under 90 seconds in length. Although our evaluation focuses on a single episode, the system is capable of processing extended podcast archives comprising many hours of audio.

To evaluate the system's ability to handle real-world questions, we curated 11 organic queries from a Tokyo Weekender article titled "11 Questions With Anime Podcast Trash Taste"[1] and a live Trash Taste QA session [2]. To expand the test set, we generated 205 synthetic queries using *GPT-4o* and 294 using *GPT-o1*. From these, we randomly sampled 50 non-duplicative synthetic queries for a final test set of 50 diverse, high-variance questions.

All text queries were then read aloud by the same male native English speaker in a controlled environment and recorded using Audacity.

3.2 Retrieval Quality

We evaluated retrieval performance with Recall@10 and normalized Discounted Cumulative Gain at 10 (nDCG@10). Following the RAGElo evaluation toolkit (Rackauckas et al., 2024) (see Appendix A), we conducted two separate evaluations using *LLM-as-a-judge* annotations, one where segments were labeled as either very relevant (1) or not relevant (0), and another where segments were labeled as somewhat relevant (1) or not relevant (0). This allowed us to assess precise retrieval performance and broader topical alignment.

Table 1: Retrieval performance of VoxRAG using cosine similarity (with and without cross-encoder–CE–reranking) on very relevant (VR) and somewhat relevant (SR) documents.

Setup	Recall@10	nDCG@10
Cosine (VR)	0.34	0.03
Cosine (SR)	**0.60**	**0.27**
Cos + CE (VR)	0.26	0.03
Cos + CE (SR)	0.46	0.14

As shown in Table 1, cosine similarity with and without a reranker retrieves segments with modest absolute scores, consistent with the challenges of speech-to-speech retrieval. Based on the large improvement in somewhat relevant over very relevant documents, while the system often retrieves topically aligned audio, it struggles to consistently retrieve precise, direct answers. The large gap between very relevant and somewhat relevant retrieval scores underscores the difficulty of fine-grained semantic matching using current audio embeddings.

3.3 Answer Quality

We evaluated answer quality along four dimensions: relevance, accuracy, completeness, and precision, each rated on a 0–2 linear scale by an impartial LLM judge using *GPT-4o*. This process follows the method of RAGElo (Rackauckas et al., 2024) (see Appendix A). Each generated answer was based on the transcripts of the top ten retrieved audio segments, along with adjacent context for continuity. Table 2 presents the mean scores and standard deviations for each evaluation dimension.

Relevance had the highest mean score (0.84), significantly outperforming all other dimensions ($p < 0.01$), with a medium-to-large effect size (Cohen's $d = 0.67$) when compared to precision. This suggests that, while the system frequently retrieved content that was topically appropriate, it often failed to deliver factual specificity or grounding. In other words, the answers were often "about the right thing," but lacked detail.

Completeness and accuracy were closely aligned, with means of 0.56 and 0.58 respectively ($p = 0.32$, $d = 0.14$), implying that partially correct answers were also seen as incomplete. Precision received the lowest average score (0.46) and was significantly lower than all other metrics. The model often failed to refer to the correct episode, moment, or speaker with sufficient granularity.

A correlation analysis reinforced these findings. Accuracy, completeness, and precision were all tightly linked ($r > 0.91$), suggesting that they capture a shared dimension of factual correctness and detail. Relevance, by contrast, was more loosely correlated with the others ($r \approx 0.77$), supporting the idea that being on-topic alone is insufficient for generating high-quality responses.

Despite the general trend, ten queries, or 20% of all queries, achieved perfect scores across all dimensions, suggesting that when embedding alignment and segment selection succeed, VoxRAG delivers strong results. A comparatively large number of queries containing the word "shower" received perfect scores, though this was not consistent across all such queries. Of the ten queries containing "shower," four achieved perfect marks, as compared to 20% overall. While anecdotal, this

[1] https://www.tokyoweekender.com/tw-community/trash-taste-podcast/

[2] https://youtu.be/tzFLreIzB78?si=yI96MWYgvQdmspsl

Table 2: Mean answer quality scores (0–2 scale) from impartial LLM judges across evaluation dimensions. Relevance significantly outperforms all other metrics. Effect sizes (d) and p-values are computed relative to relevance using two-tailed paired t-tests.

Metric	Mean	Std Dev	Δ vs. Relevance	d	p-value	Significantly Lower
Relevance	0.84	0.87	—	—	—	—
Accuracy	0.58	0.81	-0.26	0.49	< 0.01	Yes
Completeness	0.56	0.81	-0.28	0.52	< 0.01	Yes
Precision	0.46	0.81	-0.38	0.67	< 0.01	Yes

partial pattern may still reflect idiosyncrasies in how CLAP embeddings handle certain personal or lifestyle-related concepts potentially influenced by consistent acoustic or contextual cues in the training data.

4 Discussion

Our results highlight the challenges inherent in speech-to-speech retrieval within unstructured, multi-speaker podcast content. Although CLAP embeddings provided a degree of coarse semantic alignment, the retrieval process often prioritized topically related segments over precise matches. This tendency resulted in lower precision and incomplete responses. Evaluations revealed strong correlations between accuracy, completeness, and precision, indicating a shared reliance on fine-grained factual grounding. In contrast, relevance scores remained consistently high, suggesting that topical alignment alone is insufficient for generating high-quality answers. Certain queries, particularly those involving lifestyle concepts such as "shower," achieved perfect scores in some cases, but not reliably. This inconsistency may reflect variability in how well specific topics are represented within audio embeddings and warrants further investigation.

The modular architecture of VoxRAG enabled rapid experimentation across embedding models, chunking strategies, and retrieval logic. The inclusion of audio playback within the interface proved valuable for error analysis, as it revealed retrieval mismatches that were not apparent from text alone.

These findings establish a baseline for future research on audio-native question answering. They point to the need for improved embedding fine-tuning, more effective segmentation methods, and reranking strategies that better reflect factual precision. VoxRAG represents a step toward multimodal RAG systems capable of operating directly on real-world, noisy, and informal spoken content. With the proliferation of audio media, systems of this kind will be increasingly important for enabling direct retrieval and reasoning over speech, without dependence on textual transcripts.

5 Conclusion

VoxRAG explores the viability of a fully speech-to-speech retrieval pipeline for retrieval-augmented generation. While the system ultimately produces text answers, it retrieves documents directly from audio using CLAP embeddings (Elizalde et al., 2022), bypassing early transcription. Despite its novel architecture, the system underperforms on precision, completeness, and accuracy metrics, highlighting the limitations of current audio embedding models for fine-grained semantic retrieval. However, the retrieval quality on certain queries demonstrates the ultimate viability of RAG with speech-to-speech retrieval.

Limitations

While VoxRAG shows that transcription-free audio-to-audio retrieval is feasible, several challenges remain. One key limitation is the absence of transcript-based or hybrid retrieval baselines. We do not compare against methods like CLAP with transcribed input or strong text retrievers such as BM25, which makes it difficult to assess the true tradeoffs of avoiding transcription. Another issue lies in the hybrid nature of the pipeline. While retrieval is audio-only, the system still relies on Whisper transcripts for answer generation, reintroducing ASR noise and undercutting the goal of being fully transcription-free. Future work should explore audio-native generation methods.

We also note potential bias in evaluation. GPT-4o is used for both generating and assessing answers, which may lead to overestimation of performance due to model self-agreement. Using a different model, such as Qwen or Mistral, for evaluation could help mitigate this. Our evaluation is further limited by the use of only one episode from the Trash Taste podcast, restricting diversity and generalizability. Broader testing across multiple episodes and speakers would provide stronger insights.

Finally, the system shows a gap between topical relevance and factual precision. CLAP embeddings retrieve on-topic segments, but these often lack the detailed grounding needed for accurate answers. Improving fine-grained alignment remains an open challenge. These limitations are expected at this early stage and help clarify where future work can focus to strengthen audio-native retrieval and QA pipelines.

References

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Preprint*, arXiv:2006.11477.

Ann Clifton, Sravana Reddy, Yongze Yu, Aasish Pappu, Rezvaneh Rezapour, Hamed Bonab, Maria Eskevich, Gareth Jones, Jussi Karlgren, Ben Carterette, and Rosie Jones. 2020. 100,000 podcasts: A spoken English document corpus. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5903–5917, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2025. The faiss library. *Preprint*, arXiv:2401.08281.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. 2022. Clap: Learning audio concepts from natural language supervision. *Preprint*, arXiv:2206.04769.

Rosie Jones, Hamed Zamani, Markus Schedl, Ching-Wei Chen, Sravana Reddy, Ann Clifton, Jussi Karlgren, Helia Hashemi, Aasish Pappu, Zahra Nazari, Longqi Yang, Oguz Semerci, Hugues Bouchard, and Ben Carterette. 2021. Current challenges and future directions in podcast information access. *Preprint*, arXiv:2106.09227.

Oleksii Kuchaiev, Jason Li, Huyen Nguyen, Oleksii Hrinchuk, Ryan Leary, Boris Ginsburg, Samuel Kriman, Stanislav Beliaev, Vitaly Lavrukhin, Jack Cook, Patrice Castonguay, Mariya Popova, Jocelyn Huang, and Jonathan M. Cohen. 2019. Nemo: a toolkit for building ai applications using neural modules. *Preprint*, arXiv:1909.09577.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *Preprint*, arXiv:1910.13461.

Chyi-Jiunn Lin, Guan-Ting Lin, Yung-Sung Chuang, Wei-Lun Wu, Shang-Wen Li, Abdelrahman Mohamed, Hung yi Lee, and Lin shan Lee. 2024. Speechdpr: End-to-end spoken passage retrieval for open-domain spoken question answering. *Preprint*, arXiv:2401.13463.

Guan-Ting Lin, Yung-Sung Chuang, Ho-Lam Chung, Shu wen Yang, Hsuan-Jui Chen, Shuyan Dong, Shang-Wen Li, Abdelrahman Mohamed, Hung yi Lee, and Lin shan Lee. 2022. Dual: Discrete spoken unit adaptive learning for textless spoken question answering. *Preprint*, arXiv:2203.04911.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

Do June Min, Karel Mundnich, Andy Lapastora, Erfan Soltanmohammadi, Srikanth Ronanki, and Kyu Han. 2025. Speech retrieval-augmented generation without automatic speech recognition. *Preprint*, arXiv:2412.16500.

Eliya Nachmani, Alon Levkovitch, Roy Hirsch, Julian Salazar, Chulayuth Asawaroengchai, Soroosh Mariooryad, Ehud Rivlin, RJ Skerry-Ryan, and Michelle Tadmor Ramanovich. 2024. Spoken question answering and speech continuation using spectrogram-powered llm. *Preprint*, arXiv:2305.15255.

Zackary Rackauckas. 2024. Rag-fusion: A new take on retrieval augmented generation. *International Journal on Natural Language Computing*, 13(1):37–47.

Zackary Rackauckas, Arthur Câmara, and Jakub Zavrel. 2024. Evaluating rag-fusion with ragelo: an automated elo-based framework. *Preprint*, arXiv:2406.14783.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *Preprint*, arXiv:2212.04356.

Aaqib Saeed, David Grangier, and Neil Zeghidour. 2020. Contrastive learning of general-purpose audio representations. *Preprint*, arXiv:2010.10915.

Jakob Schwerter. 2022. Audio- and text-based podcast retrieval and summarization. Master's thesis, Universität Leipzig.

Silero Team. 2024. Silero vad: pre-trained enterprise-grade voice activity detector (vad), number detector and language classifier. `https://github.com/snakers4/silero-vad`. Pre-trained enterprise-grade Voice Activity Detector (VAD), Number Detector and Language Classifier.

Chunyu Sun, Bingyu Liu, Zhichao Cui, Anbin Qi, Tian hao Zhang, Dinghao Zhou, and Lewei Lu. 2025. Seal: Speech embedding alignment learning for speech large language model with retrieval-augmented generation. *Preprint*, arXiv:2502.02603.

SYSTRAN. Faster whisper. https://github.com/SYSTRAN/faster-whisper. Faster Whisper transcription with CTranslate2.

A Evaluator Prompts

A.1 Retrieval Evaluator

We used the following system prompt for our retrieval evaluator for very relevant documents:

```
You are an expert annotator evaluating
    whether a *spoken podcast transcript
     segment* is *very relevant* to a
    user's question.
These transcripts may include humor,
    casual speech, tangents, or non-
    traditional structure.
Return **1** if the segment contains
    strong, clear, and direct
    information addressing the user's
    question.
Return **0** if the segment is only
    loosely or partially related, or
    entirely off-topic.
Only respond with a single digit: 1 or
    0. Do not explain.
```

For evaluating somewhat relevant documents, we used the following prompt:

```
You are an expert annotator evaluating
    whether a *spoken podcast transcript
     segment* is *somewhat relevant* to
    a user's question.
These transcripts may include humor,
    casual speech, tangents, or non-
    traditional structure.
Return **1** if the segment has a loose
    or minor connection to the user's
    question - it may touch on a related
     theme, mention something adjacent,
    or vaguely resemble the topic, even
    if it is incomplete or off-target.
Return **0** if the segment has no real
    connection at all.
Only respond with a single digit: 1 or
    0. Do not explain.
```

A.2 Answer Evaluators

For the answer quality evaluation, we used the following prompt:

```
You are an impartial judge for
    evaluating the quality of the
    responses provided by an AI
    assistant tasked with answering
    users' questions about the *Trash
    Taste* podcast.

You will be given the user's question
    and the answer produced by the
    assistant. The assistant's answer
    was generated based on a set of
    audio-derived documents retrieved
    from episodes of the *Trash Taste*
    podcast.

You will be provided with the relevant
    podcast segments retrieved by the
    search engine.

Your task is to evaluate the answer's
    quality based on the response's **
    relevance**, **accuracy**, **
    completeness**, and **precision**,
    grounded in the retrieved podcast
    content.

## Rules for evaluating an answer:
- **Relevance**: Does the answer address
     the user's question?
- **Accuracy**: Is the answer factually
    correct, based on the retrieved
    podcast segments?
- **Completeness**: Does the answer
    provide all the information needed
    to address the user's question?
- **Precision**: If the user asks about
    a specific episode, moment, guest,
    or topic, does the answer correctly
    identify and reflect that specific
    context?

## Steps to evaluate an answer:
1. **Understand the user's intent**:
    Restate what the user is trying to
    find out, in your own words.
2. **Check if the answer is correct**:
    Think step-by-step about whether the
     answer truthfully and fully
    responds to the user's question.
3. **Evaluate the quality of the answer
    **: Judge the answer on relevance,
    factual accuracy (according to the
    retrieved podcast segments), and how
     completely it covers the query.
4. **Assign a score**: Produce a single-
    line JSON object with the following
    keys, each with a score from 0 to 2:

- "relevance"
  - 0: The answer is not relevant to the
      user's question.
  - 1: The answer is partially relevant.
  - 2: The answer is fully relevant.

- "accuracy"
  - 0: The answer is factually incorrect
     or contradicts the retrieved
     content.
  - 1: The answer is partially correct
     but includes errors or
     misinterpretations.
  - 2: The answer is factually correct
     based on the retrieved segments.

- "completeness"
  - 0: The answer leaves out major parts
      of the question.
  - 1: The answer addresses the question
      only in part.
  - 2: The answer covers all key aspects
      of the user's question.

- "precision"
```

```
    - 0: The answer refers to the wrong
        episode, topic, or context.
    - 1: The answer is somewhat related
        but not specific enough.
    - 2: The answer directly reflects the
        specific content or moment asked
        about.

The last line of your evaluation must be
     a SINGLE LINE JSON object with the
    keys "relevance", "accuracy", "
    completeness", and "precision", each
     assigned a score between 0 and 2.

[ DOCUMENTS RETRIEVED ]
{documents}

[ User Query ]
{query}

[ Agent Answer ]
{answer}
```

B Hardware and Runtime Configuration

Experiments were conducted on two machines: one with an RTX 5090 GPU, and another with two NVIDIA A4500 GPU's. On the latter machine, from user input to LLM output, the average generation time was 3.02 seconds. Under these conditions, the system demonstrates practical viability for offline indexing and interactive, low-latency query-time retrieval.

C Embedding Hyperparameters

Audio segments and spoken queries were embedded using the *laion/clap-htsat-unfused* model. Prior to indexing and similarity search, both query and document embeddings were L2-normalized. This normalization step ensures that inner product similarity in the FAISS `FlatIP` index is equivalent to cosine similarity. This enables semantically aligned retrieval directly in the CLAP embedding space without requiring additional reranking or feature engineering.

D Sample QA Pairs

Table 3: A sample of QA pairs on Synthetic Queries

QA pairs (Q = white, A = blue)
Q: Can the one-minute shower routine ever replace your normal shower, or is that impossible?
A: ... Connor expresses skepticism about whether such a quick shower can truly replace a normal shower. He mentions that taking a shower is not just about getting clean but also about the time it takes to refresh and wake up properly. Joey adds that a shower under five minutes might not be sufficient to adequately clean oneself, suggesting that such a quick routine is more about covering up body odor than truly cleaning. Essentially, both seem to agree that while a one-minute shower might serve in a pinch, it cannot fully replace the benefits of a longer, more thorough shower. This perspective is captured in Segment 17.
Q: What travel conditions now make or break a trip for you?
A: ... Garnt shared that he finds Kyoto to be a very touristy spot, which can detract from the experience. He mentioned that while it's worth visiting at least once, the crowds and the commercialization of the area can take away from its magic. He expressed a preference for the outskirts of Kyoto, which tend to be more enjoyable and less crowded (Segment 20). Connor added that the disparity between Kyoto's historical significance and its touristy nature can make visits feel less authentic. He indicated that while it might be great for first-time visitors, the experience might not be as enjoyable upon return due to the overwhelming number of tourists (Segment 20). Overall, both speakers highlight that the level of crowding and commercialization in popular tourist destinations can make or break the quality of a travel experience.
Q: If you made a Trash Taste Bingo Card, what squares would be absolutely required?
A: ... 1. Mention of League of Legends — This game often comes up in discussions (Segment 2). 2. Reference to drinking stories — The guys frequently share hilarious and embarrassing drinking experiences (Segment 5). 3. Discussion about sibling dynamics — They talk about their experiences growing up as siblings or only children (Segment 26). 4. Anime recommendations — Each episode usually involves recommendations or debates about anime (Segment 19). 5. Joey's love for food — Food discussions, especially around Japanese cuisine, are a staple (not specifically cited, but a recurring theme). 6. Mention of gaming frustrations — Conversations about the challenges of gaming, especially in competitive environments (Segment 17). 7. Cultural references — They often reference their cultural backgrounds, especially regarding Australia and the UK (Segment 10) ...