

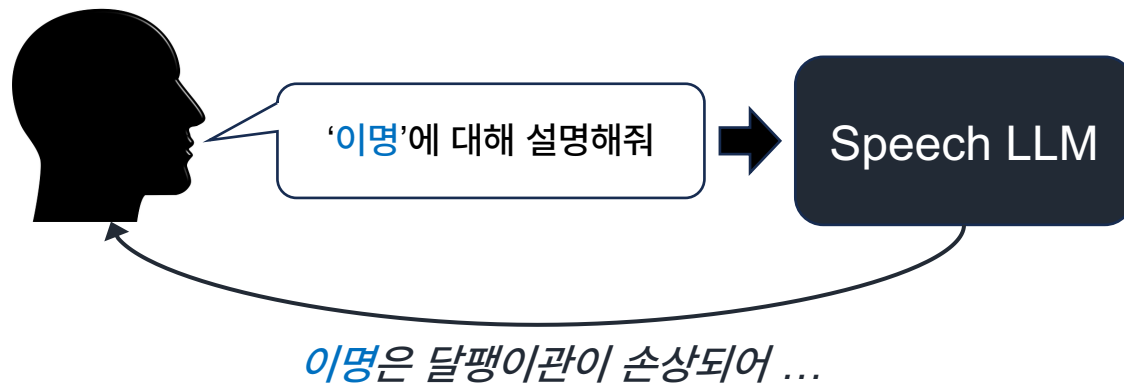
Lightweight SpeechLLM for Medical Speech Recognition and QA

김종하

2025. 9. 7

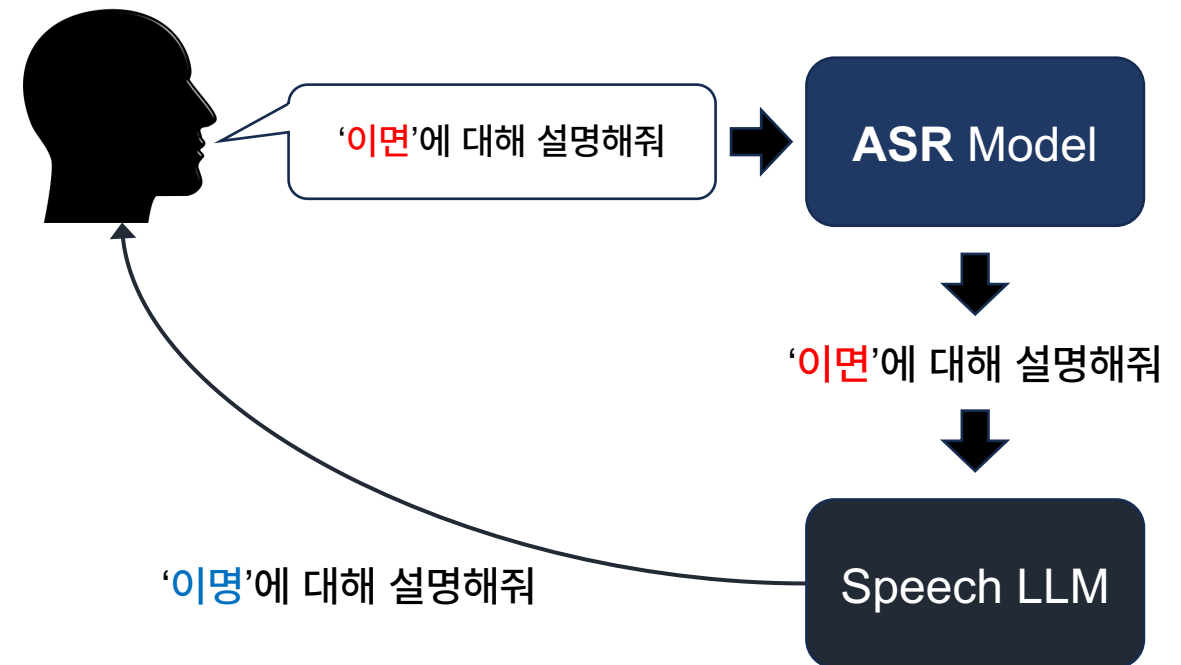
Field Definition

- Speech LLM
 - 음성/오디오를 직접 이해하고 답변할 수 있는 멀티모달(multi-modal) 대규모 언어 모델
 - 구성
 - Audio Encoder
 - Modality Adapter
 - Pretrained LLM



• LLM-based ASR Error Correction

- 전사된 음성을 언어 모델을 이용해 수정하는 방법
- Phonetic / Semantic / Grammatical 오류 교정



Results Overview

Table 1: ASR (Key Performance Metrics by ASR Model)

Model	Latency (s)	RPS	RTF	Model size (GB)	CER (%)
voxtral 3B 4bit	1.21	0.83	5.86	2.18	19.14
whisper large FP16	0.75	1.32	7.74	2.87	8.66
ours	1.55	0.65	3.95	2.25	9.09

• ASR

- Latency : 오디오 샘플 1개를 처리하는 데 걸리는 평균 시간(초)
- RPS : 1초당 처리할 수 있는 오디오 샘플의 개수
- RTF : 실제 오디오 길이 대비 처리 속도
- CER : 변환된 텍스트가 정답과 문자 단위 오류율

Table 2: SQA(Speech Question Answering) GPT-OSS Model evaluation

Evaluation Model	Average Score (out of 5 points)
Voxtral 3B 4-bit (base)	2.07
Voxtral 3B 4-bit + RAG	2.91
Voxtral 3B 4-bit + LoRA (Ours)	4.01

• SQA

• LLM-as-a-Judge (GPT-OSS)

- ▶ 5점 (매우 우수): 질문의 핵심 의도를 완벽하게 이해하고, 매우 정확하고 상세하며 직접적인 답변을 제공함.
- ▶ 4점 (우수): 질문에 직접적으로 답변하며, 대부분의 내용이 정확하고 관련성이 높음. 약간의 아쉬움이 있을 수 있음.
- ▶ 3점 (보통): 질문의 일부에만 답변하거나, 답변의 관련성이 다소 떨어짐.
- ▶ 2점 (미흡): 질문의 핵심을 파악하지 못하고, 관련 없는 정보를 제공하거나 답변에 오류가 많음.
- ▶ 1점 (매우 미흡): 질문과 전혀 관련 없는 답변을 하거나, 완전히 잘못된 정보를 제공함.

Research Background and Problem Definition

- 순천향대 Task 정의
 - 특정 문맥을 위한 실시간 음성 인식 시스템 구축
 - 이비인후과 진료 과정 중 의료 용어에 대한 인식률 향상이 목표
- 기존 접근법 (audio contextual biasing)
 - Domain Adaptation
 - 의료 음성 데이터로 모델을 재학습
 - 현재 도메인 음성 데이터가 부족, Annotation 비용 소모
 - Vocabulary Injection
 - 의료 사전이나 용어 리스트를 삽입해 인식률 보정
 - 기본 ASR 성능에 강하게 의존, vocab 규모 커지면 혼선/효율 저하
 - LM Rescoring
 - KENLM, RNNLM
 - 후처리 언어모델로 후보 문장 확률 재평가
 - 몇개의 앞뒤 단어로 추론하기 때문에 의미 이해, 임상적 추론 불가능

The Need for SpeechLM / Audio Multimodal LLM

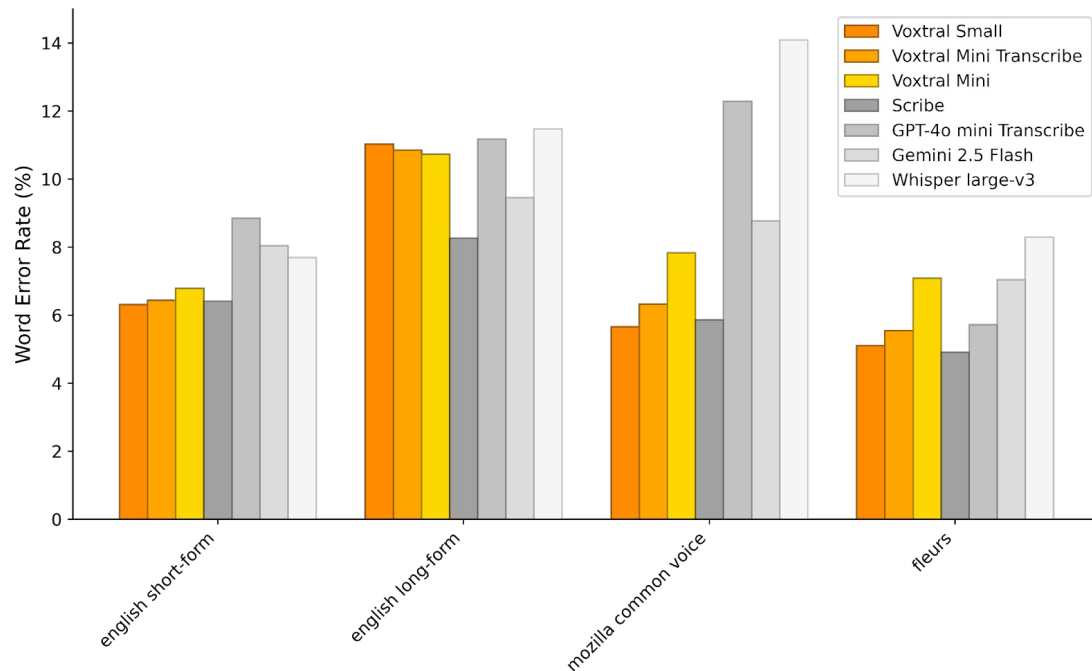
- 오디오 임베딩 + LLM 디코더를 통해 단순 전사를 넘어 맥락적 이해·응답 가능
 - 잘 듣는 것(ASR) → 잘 이해하고 답하는 것(SpeechLM)
 - 사람이 이해하는 방식과 가장 유사하기 때문에 앞으로 ASR 분야가 나아가야 할 방향
- 음성 데이터 보다 텍스트 데이터의 양과 퀄리티가 더 뛰어나기 때문에, 특수 도메인 적응 속도/효율 향상
- 결과적으로, 의료 음성을 “듣고 이해하고 답하는 모델”로 발전시킬 수 있음

Research Objectives

- 가설
 - LLM decoder가 가진 의료 지식이 ASR 성능을 향상 시킬 것이다.
- 목표
 - 의료 진료 상황에서 의료용어 **ASR 성능 향상**
 - 의학 지식 Speech QA 성능 향상
- 조건
 - **QA task** 를 수행하여 모델이 의료 지식 수준을 판단.
 - On-device, 개인정보 보호
 - 4-bit 양자화로 VRAM 절감
 - 실시간성 확보 (RTF↓, RPS↑)
- 실험 내용
 1. RAG
 2. LLM finetuning
 - TQA data
 3. LLM + audio adapter finetuning
 - Stage1 : TQA data
 - Stage2 : SQA, ASR data

Select Backbone

- Mistralai의 Voxtral mini 3B
 - LLM based E2E Transcription
 - Speech QA, Text QA
 - Open-Source
 - Simple Structure



Transcription mode



QA mode

Select Backbone

• Mistralai의 Voxtral mini 3B 문제점

- 의료 질문에 높은 Hallucination
- 낮은 의료 단어 인식률
- 한국어 미지원 (추가 학습 없음)
 - Whisper, LLM 자체 한국어 성능에 의존
- 문어체 형식 모델 답변

Table 2: 성능 비교 (한국어)

발화 종류	FP16(%)	4bit(%)
일반 발화 (ksponspeech)	32.58	40.22
의료 발화 (이비인후과 진료)	19.14	21.18

• Mistralai의 Voxtral mini 3B 특징

- 답변을 수정
 - 정답: 우린 이제 다 익혀 **먹잖아**. **으지 그런 그러다 보니까 이제** 부드러운 고기만 먹다 보니까 이제 쓸 필요가 **없어가지고** 사랑니가 이제 이상하게 나기 시작하는 **거거든**.
 - 예측: 우린 이제 다 익혀 **먹죠?** **그래서** 부드러운 고기만 먹다 보니 이제 쓸 필요 **없어서** 사랑니가 이상하게 나기 시작하는 **거죠**.

출처 : Ksponspeech

Dataset

- ASR dataset
 - AI Hub 한국어 비대면 진료(의료진 음성 33,000 문장(700시간), 환자 45만 문장(600시간))
- TQA dataset (Text Question & Answer)
 - AI Hub 초거대 헬스케어 질의응답(240,454개)
 - 아산 AMC(19,020개)
- SQA dataset (Audio Question, Text Answer)
 - AI Hub 초거대 헬스케어 질의응답 이비인후과 과목(16401개)
 - TTS로 음성 생성

Dataset Problem Solving

- TQA dataset (Text Question & Answer)
 - 초거대 헬스케어 질의응답(195,001개)
 - ▶ 질문과 정답의 pair가 불명확한 문제
 - ▶ GPT-oss model 20B에 정답을 넣고 질문 생성
 - ▶ 중복되거나, 불필요한 답변 제거
- SQA dataset (Audio Question, Text Answer)
 - 초거대 헬스케어 질의응답 이비인후과 과목 한정
 - ▶ Cosyvoice 0.5B 모델 사용
 - ▶ GPT-oss model로 생성한 질문을 변환
 - ▶ Whisper large v3 모델 CER 평가 – 필터링

```
{
  "role": "system",
  "content": """"당신은 주어진 의료 답변을 보고, 그 답변을 이끌어냈을 환자의 원래 질문을 역으로 추론하는 AI입니다.
다음 규칙을 반드시 준수하세요:
1. **반드시 한국어로만, 그리고 오직 하나의 질문만** 생성하세요.
2. 질문은 실제 환자가 말하는 것처럼 **자연스러운 대화체**여야 합니다.
3. 번호, 목록, 부연 설명 등 **질문 외의 다른 텍스트는 절대 포함하지 마세요.**""""
},
{
  "role": "user",
  "content": f""""다음은 '{department}' 진료과목의 '{disease_name}'에 대한 답변입니다. 이 답변이 나오게 된 원래 질문을 생성해 주세요.
[답변]
{answer_text}

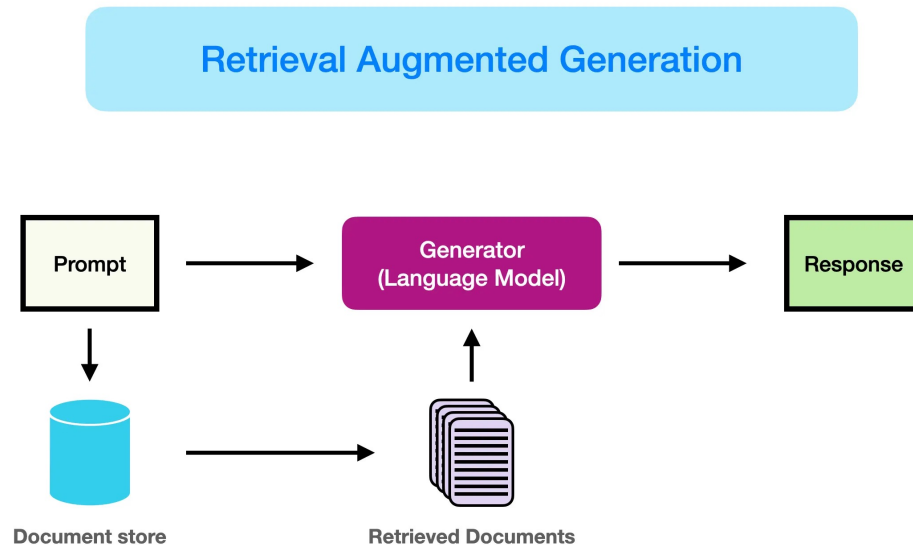
[생성할 질문]""""
}
```

입력 Prompt

Title of Slide - 슬라이드 제목

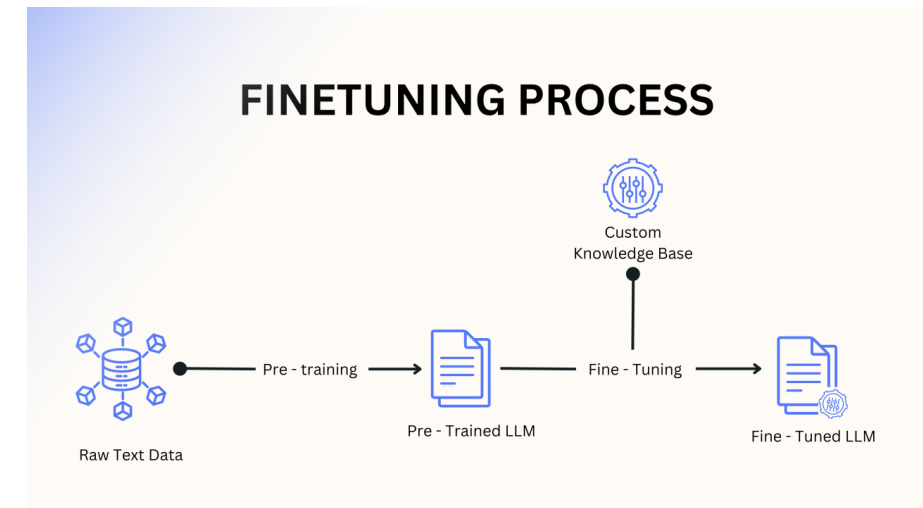
• RAG (Retrieval-Augmented Generation)

- 데이터베이스나 문서 집합에서 관련 정보를 검색하고 검색한 정보를 바탕으로 텍스트를 생성하는 방법
- 별도의 학습이 필요하지 않음



• Finetuning

- 사전 학습된 기존 언어 모델에 추가로 외부 데이터를 더 학습시키는 방법
 - Full-finetuning
 - Knowledge Distillation
 - LoRA
- 시간과 비용이 소모



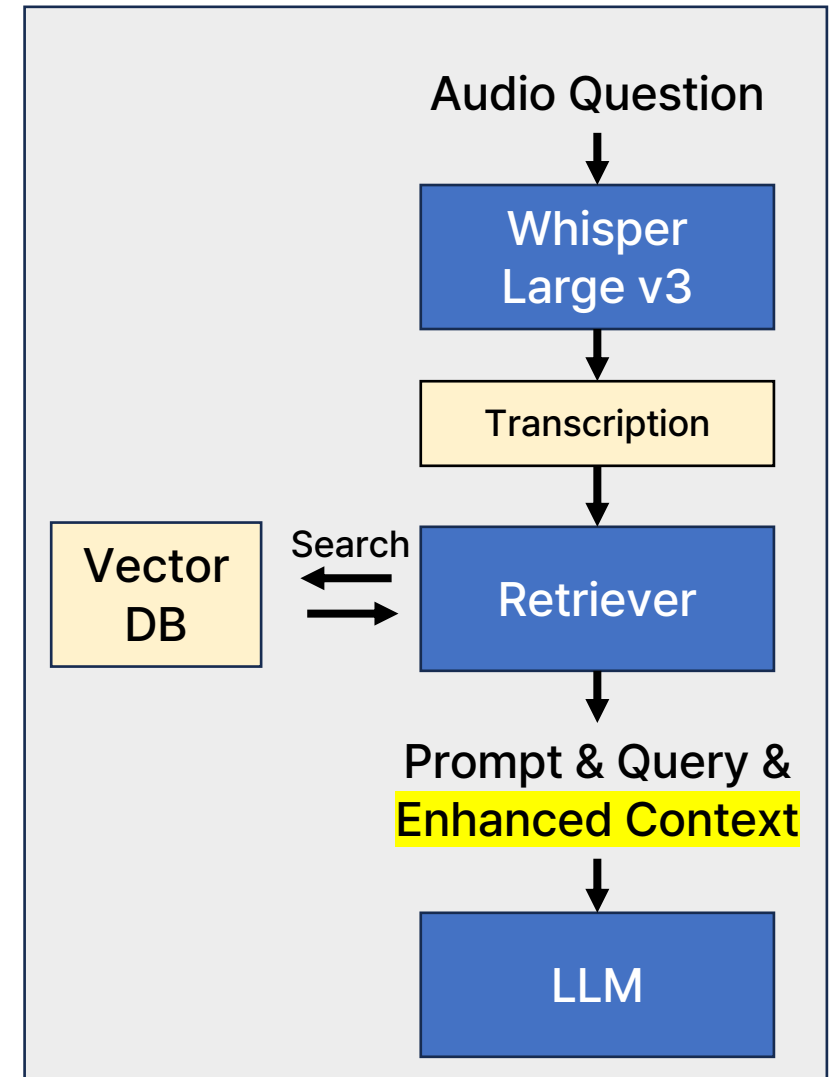
Training Strategies : RAG

• 방법

- 의학 정보를 벡터화 하여 DB로 저장
- 전사된 문장과 가장 유사도가 높은 데이터를 언어모델에 제공
- 참조
 - [1] C. Xiao, Z. Hou, D. Garcia-Romero, and K. Han, "Contextual ASR with retrieval augmented large language model," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2025.
 - [2] S. Li et al., "LA-RAG: Enhancing LLM-based ASR accuracy with retrieval-augmented generation

• 데이터셋

- AIHub 필수 의료 의학지식, 전문 의학 지식, 초거대 헬스케어 QA 데이터 사용



Model structure

Training Strategies : RAG

- 결과
 - 답변이 영어나 일본어로 나오는 Language Confusion이 줄어듦
 - Language Confusion 무시하면 2.92, 3.01 (base, RAG)
 - ASR 오류가 Retriever에 성능 저하 요인이 됨
 - 예시 : 턱관절장애 - 척추 관절 장애

Table 1: SQA(Speech Question Answering) GPT-OSS Model evaluation

Evaluation Model	Average Score (out of 5 points)
Voxtral 3B	2.07
Voxtral 3B, RAG	2.9

Training Strategies : RAG

• 전

- 나의 귀가 이상이 있는 것 같아요. 왜냐하면 소리를 구별하지 못해요. 혹시 이물질이 들어갔거나 감염이 있었나요?

あなたが感じている症状は、耳の機能に問題があることを示唆しています。具体的には、音の識別能力が低下している可能性があります。以下にいくつかの原因と対策を挙げます。### 可能性のある原因 1. **外耳道の異物**:- 耳に異物が入った場合、音が聞こえにくくなる場合があります。 - **対策**:- 専門医に相談し、異物を取り除く処置を受けること。 2. **中耳炎**:- 中耳に炎症が起こると、音が聞こえにくくなる場合があります。 - **対策**:- 炎症を鎮めるための薬を処方されることが多いです。 3. **耳の感染症**:- 耳の感染症（例えば、中耳炎や外耳炎）が原因で音が聞こえにくくなる場合があります。 - **対策**:- 感染症を治療するための薬を処方されることが多いです。 4. **内耳の問題**:- 内耳の機能に問題がある場合、音の識別能力が低下することがあります。 - **対策**:- 専門医に相談し、内耳の機能を評価するための検査を受けること。 5. **高齢による聴力低下**:- 高齢になると、自然と聴力が低下することがあります。 - **対策**:- 聴力補助器を使用することが有効です。 6. **耳鳴り**:- 耳鳴りが原因で音が聞こえにくくなる場合があります。 - **対策**:- 耳鳴りの原因を特定し、適切な治療を受けること。 ### 対策 1. **専門医に相談する**:- 早期に専門医に相談し、適切な診断と治療を

• 후

- 나의 귀가 이상이 있는 것 같아요. 왜냐하면 소리를 구별하지 못해요. 혹시 이물질이 들어갔거나 감염이 있었나요?

귀에 이상이 있는 경우, 여러 가지 원인이 있을 수 있습니다. 소리를 구별하지 못하는 증상은 귀의 기능이 저하된 것을 나타낼 수 있습니다. 귀에 이물질이 들어갔거나 감염이 있었는지 확인하기 위해 전문의에게 상담하는 것이 좋습니다. 귀의 기능 저하를 일으킬 수 있는 몇 가지 원인은 다음과 같습니다:

1. ****이물질****: 귀에 먼지, 먼지, 작은 물방울, 귀털 등이 들어갈 수 있습니다. 이물질이 귀에 있으면 귀를 청소하는 것이 필요할 수 있습니다.
2. ****감염****: 귀의 감염은 귀의 기능을 저하시킬 수 있습니다. 감염은 귀의 염증, 통증, 발열, 귀가 막힌 느낌 등을 동반할 수 있습니다.
3. ****귀염****: 귀염은 귀의 염증으로, 귀의 기능을 저하시킬 수 있습니다. 귀염은 종종 감염으로 인해 발생하며, 진통제와 항균제 치료가 필요할 수 있습니다.
4. ****귀의 손상****: 외상, 내분비 질환, 중독, 약물 사용 등 다양한 요인이 귀의 기능을 저하시킬 수 있습니다.
5. ****청각 손상****: 청각 손상은 귀의 기능을 저하시킬 수 있습니다. 청각 손상은 노화, 중독, 외상, 감염 등 다양한 요인에 의해 발생할 수 있습니다.
6. ****신경계 문제****: 뇌의 신경계 문제도 귀의 기능을 저하시킬 수 있습니다.

이러한 증상을 겪고 있다면, 즉시 전문의에게 상담하는 것이 중요합니다. 전문의는 귀의 상태를 진단하고 적절한 치료를 추천할 수 있습니다. 귀의 청소, 항균제, 진통제, 수술 등 다양한 치료 방법이 있을 수 있습니다.

Training Strategies : RAG

• 전

- 어떻게 하면 고막염을 정확하게 진단할 수 있는지 알고 싶습니다.

Bomangyeom (보망염)은 한국어에서 "보병염"으로 번역될 수 있으며, 이는 주로 한국에서 사용되는 용어입니다. 정확한 진단을 위해서는 전문적인 의료 지원을 받는 것이 중요합니다. Bomangyeom의 정확한 진단은 다음과 같은 단계로 이루어질 수 있습니다:

1. **의사 상담 및 검사**:

- **의사 상담** : 의사와 상담하여 증상을 자세히 설명합니다.
- **신체 검사** : 신체 검사를 통해 증상을 확인합니다.
- **혈액 검사** : 혈액 검사를 통해 염증을 확인할 수 있습니다.

2. **이미지 검사**:

- **X선 검사** : 관절의 구조를 확인하는 데 사용됩니다.
- **CT 스캔** : 더 자세한 구조를 확인하는 데 사용됩니다.
- **MRI 스캔** : 관절의 구조와 염증을 더 자세히 확인하는 데 사용됩니다.

3. **혈액 검사**:

- **CRP 검사** : 염증을 확인하는 데 사용됩니다.
- **백혈구 계수** : 염증을 확인하는 데 사용됩니다.

4. **진단 검사**:

- **관절 점검** : 관절의 상태를 직접 확인하는 데 사용됩니다.

...

• 후

- 어떻게 하면 고막염을 정확하게 진단할 수 있는지 알고 싶습니다.

보망염(보그염, Bovine Spongiform Encephalopathy)은 소의 뇌에 영향을 미치는 전염병으로, 인간에게도 전염될 수 있는 위험이 있습니다. 보망염을 정확하게 진단하기 위해서는 여러 가지 단계가 필요합니다. 다음은 보망염 진단을 위한 주요 단계입니다:

1. **병리학적 검사**:

- **병리학적 검사** : 소의 뇌 조직을 검사하여 병리학적 변화를 확인합니다. 보망염은 뇌 조직이 가려워지는 특성을 가지고 있습니다.
- **전자 미세구조 검사 (EM)** : 뇌 조직의 전자 미세 구조를 분석하여 보망염 특성을 확인합니다.

2. **생물학적 검사**:

- **프로테아스 검사** : 보망염은 특정 종류의 프로테아스(蛋白質)를 포함합니다. 이 프로테아스를 검출하여 보망염을 진단할 수 있습니다.
- **RNA 검사** : 최근에는 RNA 검사를 통해 보망염을 진단하는 기술도 개발되었습니다.

3. **진단 검사**:

- **인체 검사** : 보망염은 인간에게도 전염될 수 있으므로, 보망염 진단을 위해 소의 뇌 조직을 사용하지 않고도 보망염을 진단할 수 있는 기술이 필요합니다. 예를 들어, 소의 혈액이나 다른 조직을 검사하여 보망염을 진단할 수 있습니다.

...

Training Strategies : RAG

- Conclusion
 - ASR + RAG는 Error propagation 때문에 틀린 답변을 선택하는 경향이 있다
 - WavRAG: Audio-Integrated Retrieval Augmented Generation for Spoken Dialogue Models [ACL 2025]
 - LLM에 RAG를 제공했을 때 환각(hallucination) 현상을 보이거나, 입력된 1차 가설의 길이를 무시하고 너무 길거나 짧은 문장을 생성하는 문제가 있다.
 - ASR 오류 교정 작업에 특화되도록 미세조정(fine-tuning) 하는 것이 성능 향상에 좋았다.
 - Contextual ASR with Retrieval Augmented Large Language Model , AWS AI Labs [ICASSP 2025]

Training Strategies : Finetuning

- LLM only finetuning

- 모델이 이미 오디오-텍스트 관계를 잘 학습했다면, Text finetuning 만으로 의료음성인식 성능이 향상될 것이다.
- Medical dataset
 - Text QA
- LLM only LoRA finetuning

- 2 stage finetuning

- 모델이 제공하는 Transcription, Speech QA 기능을 모두 의료 데이터셋으로 finetuning
- Medical dataset
 - ASR
 - Text QA (TQA)
 - Speech QA (SQA)
- LLM, Adapter LoRA finetuning

Stage 1 : LLM only finetuning

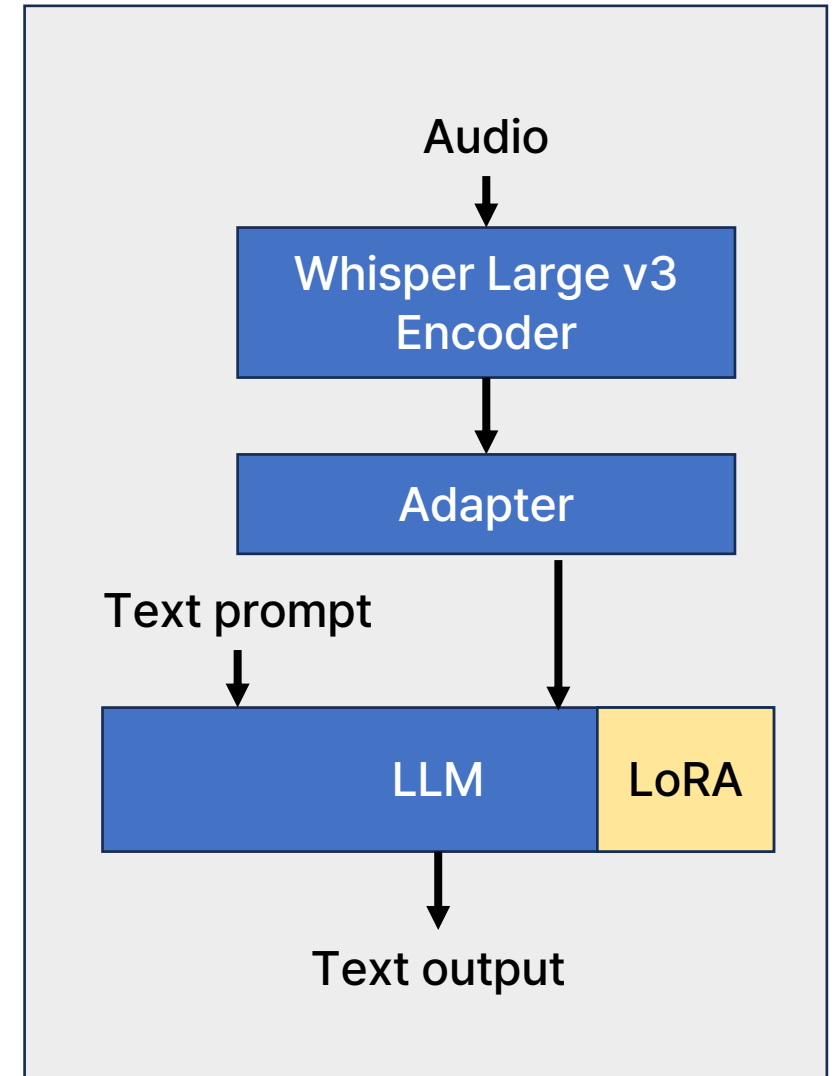
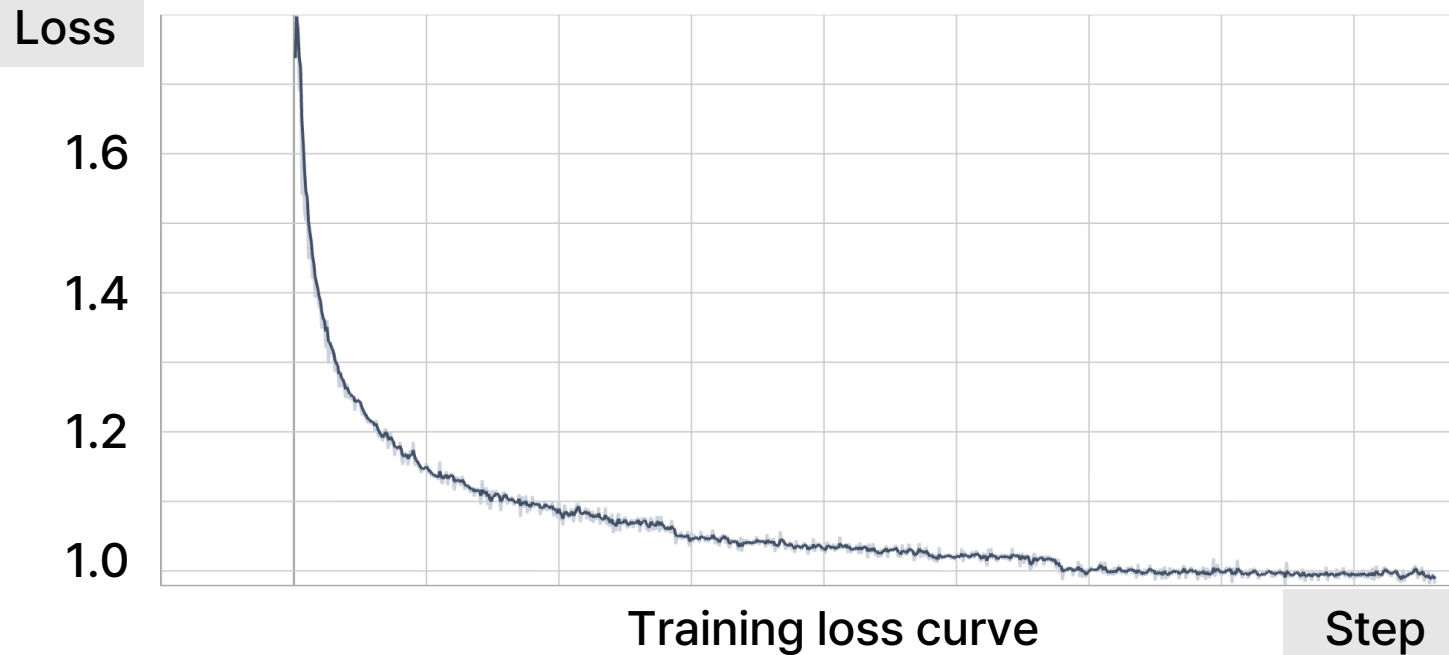
- 이전 방법에서 사용했던 모델 활용
- TQA 데이터로 의료 도메인 지식 학습

Stage 2 : LLM, Adapter LoRA finetuning

- Adapter + LLM LoRA finetuning
- ASR, SQA 데이터로 오디오-텍스트 관계 정렬

Training Strategies

- LLM only finetuning
 - 방법
 - ▶ 초거대 헬스케어 질의응답 데이터 사용
 - ▶ LLM QLoRA finetuning (Quantized Low-Rank Adaptation)
 - 16GB VLAM 최적화
 - NF4, 이중 양자화, 혼합 정밀도 계산



Model structure

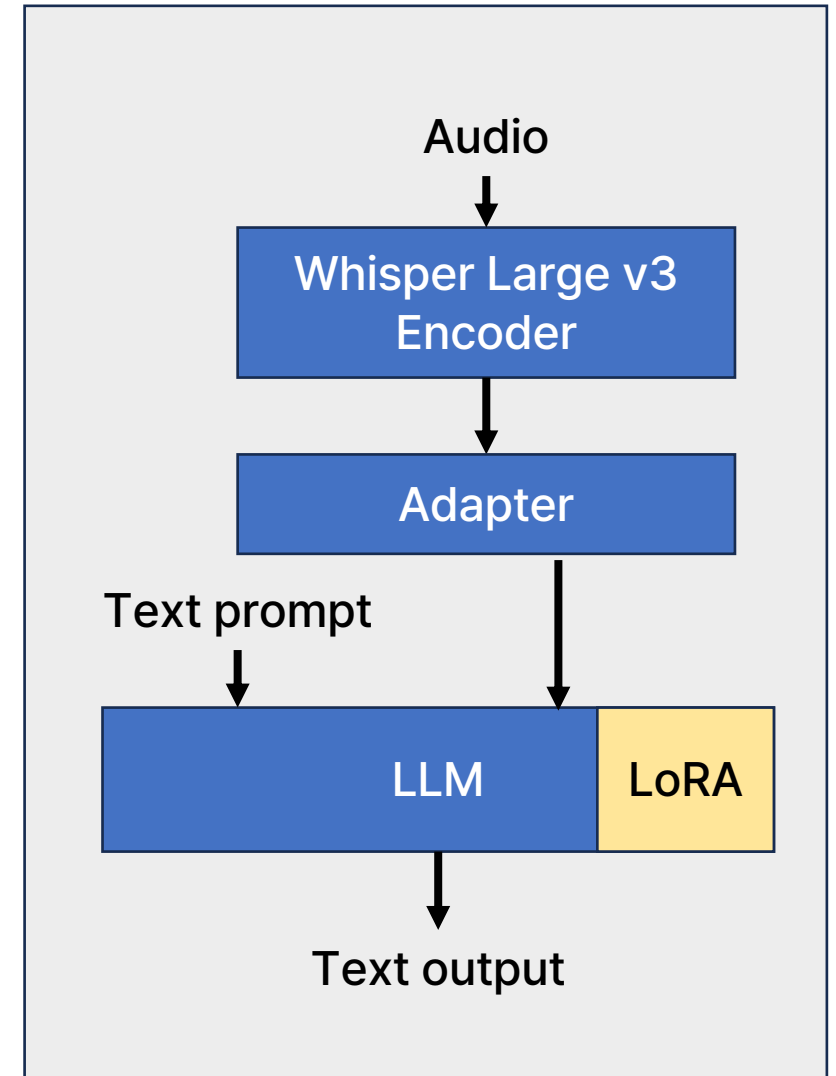
Training Strategies

- LLM only finetuning
 - 결과
 - catastrophic forgetting 발생
 - 기존 음성 Transcript 방법을 망각
 - Transcript 이 아닌 QA 응답을 반환

Transcription Task

Input : 어떻게 하면 상기도 감염을 예방할 수 있을까요?

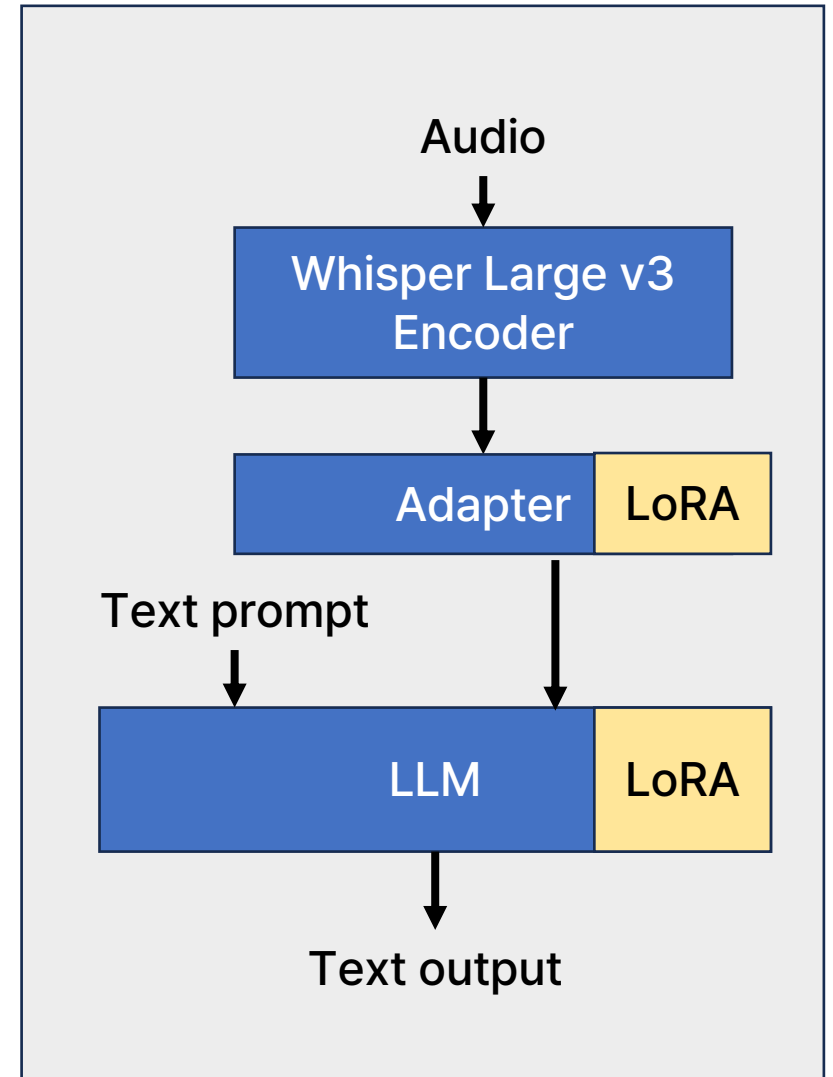
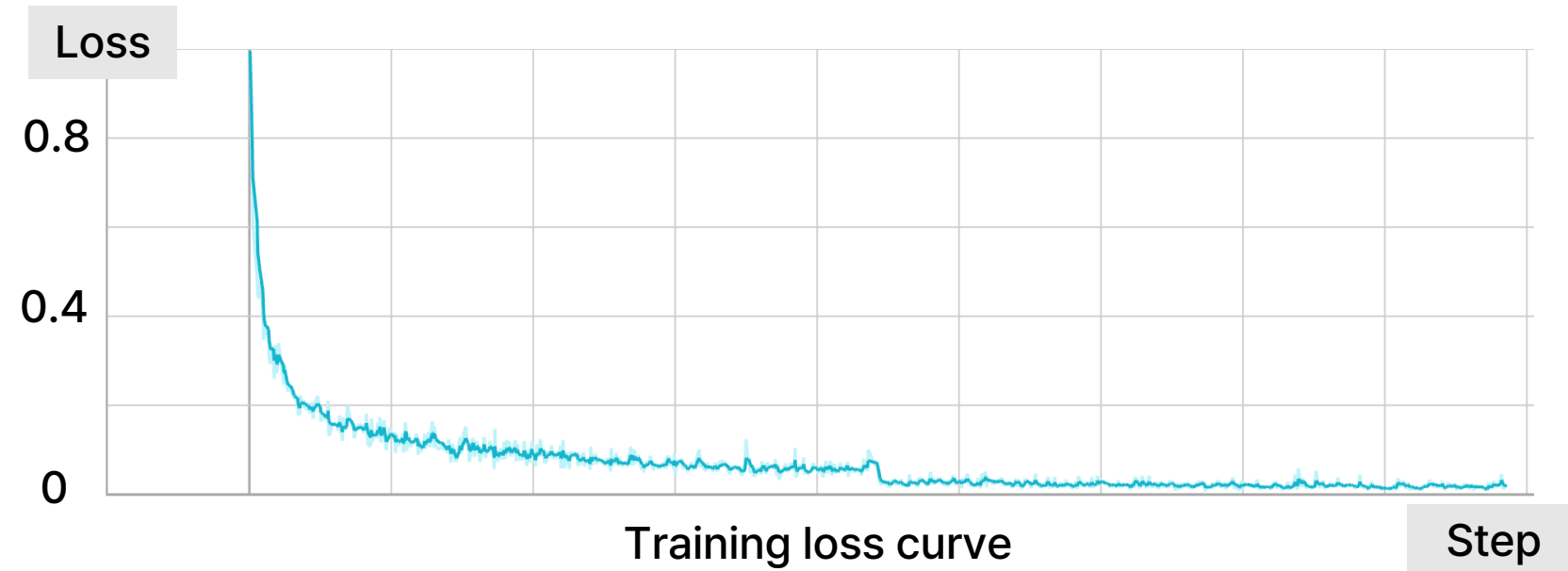
Output : 상기도 감염을 예방하기 위해서는 몇 가지 주의사항을 지켜야 합니다. 먼저 손을 깨끗이 씻는 것이 매우 중요합니다. 외출 후 화장실 사용 후 식사 전에는 반드시 손을 깨끗하게 씻어야 합니다. 또한 감기 증상이 있는 사람과의 접촉을 피해야 합니다. 증상이 지속되거나 악화 되면 의료진과 상담을 권합니다.



Model structure

Training Strategies

- 2 stage finetuning
 - 방법
 - ▶ Stage 1 : 기존에 LLM LoRA finetuning 방법 사용
 - 의료 도메인 지식 학습
 - ▶ Stage 2 : ASR, SQA 데이터로 음성 학습
 - Audio encoder Adapter + LLM LoRA finetuning



Model structure

Training Strategies

- 2 stage finetuning
 - 결과
 - ▶ 의료 ASR 성능 향상
 - ▶ ASR 데이터셋에 없는 의료 단어도 인식

Transcription

Answer : 비중격성형술 및 필요시 비중격확장이식을 고려할 수 있습니다.

Base : "비전격성형술" 및 "필요시 비전격확장이식"을 고려할 수 있습니다.

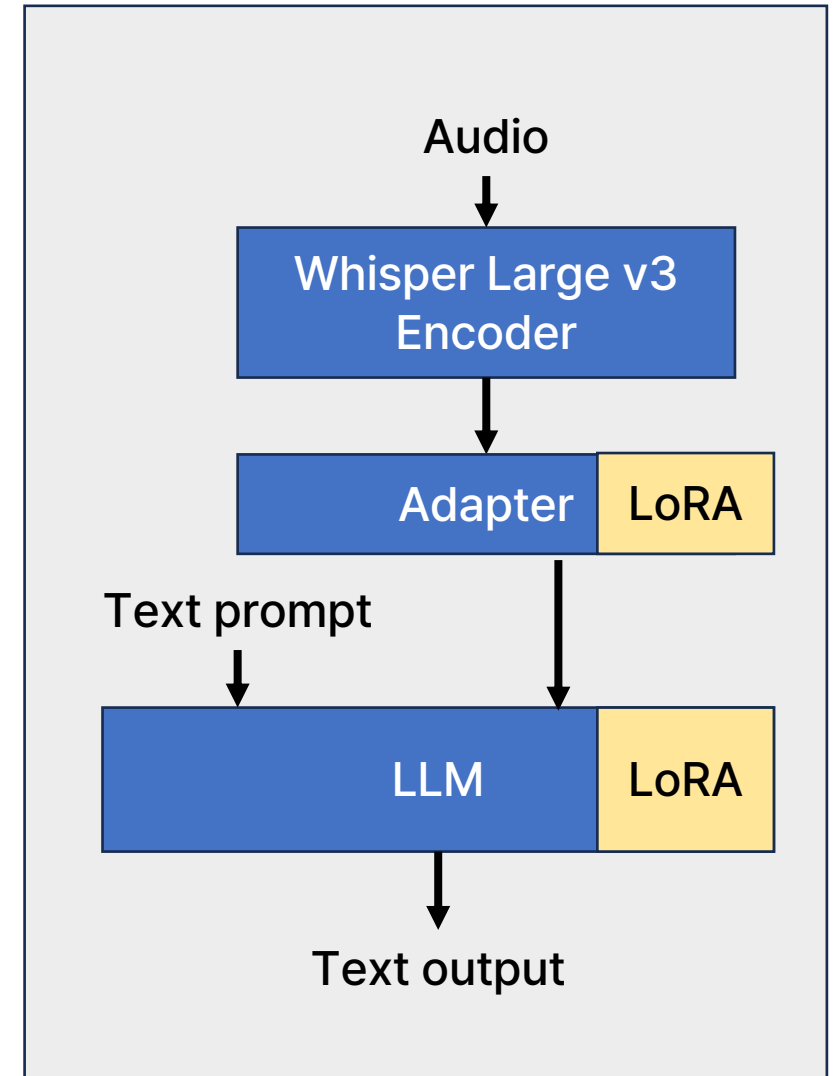
Ours : 비중격 성형술 및 필요시 비중격 확장 이식을 고려할 수 있습니다.

Answer : 전산화단층촬영으로 구조를 정밀히 보고, 필요하다면 두개저수술이나 코교정술 접근을 통해 병변을 절제해 조직검사를 합니다.

Base : 전산나 단침 촬영으로 구조를 정밀히 보고, 필요하다면 두개조 수술이나 코구정술 접근을 통해 병변을 절제해 조직 검사를 합니다.

Ours : 전산나단층촬영으로 구조를 정밀히 보고 필요하다면 두개저수술이나 코교정술 접근을 통해 병변을 절제해 구조적 검사를 합니다.

Answer : 정답 , Base : Voxtral mini 3B , Ours : Base 모델 의료 Finetuning



Model structure

Training Strategies

• 전

- 어떤 요인으로 인해 후비루가 발생할 수 있는지 알고 싶어요?

QA 이전 결과 넣기

후비루는 주로 두 가지 주요 요인에 의해 발생할 수 있습니다:

- **비타민 B12 결핍****: 비타민 B12는 신경계의 정상적인 기능과 세포 생성을 위해 필수적입니다. 비타민 B12 결핍은 후비루의 주요 원인 중 하나입니다. 비타민 B12 결핍은 여러 가지 이유로 발생할 수 있습니다:
 - **식단 부족****: 비타민 B12는 주로 고기, 계란, 우유, 치즈, 생선 등에서 공급됩니다. 비타민 B12를 충분히 섭취하지 않는 경우 결핍이 발생할 수 있습니다.
 - **소화기 문제****: 소화기 질환, 특히 소화불량이나 장염은 비타민 B12 흡수에 영향을 미칠 수 있습니다.
 - **장내 흡수 장애****: 장내 흡수 장애는 비타민 B12를 소화관에서 흡수하지 못하게 합니다. 이는 장염, 장암, 장의 염증성 질환 등으로 인해 발생할 수 있습니다.
 - **장내 흡수 장애 치료****: 일부 약물, 예를 들어 니코틴산, 수크라린, 메트로트렉산, 클로피라미드, 프로파몰, 아미오딜, 메스나, 메스나-아미오딜, 메스나-프로파몰, 메스나-프로파몰-아미오딜, 메스나-프로파몰-아미오딜-클로피라미드, 메스나-프로파몰-아미오딜-클로피라미드-프로파몰, 메스나-프로파몰-아미오딜-클로피라미드-프로파몰-아미오딜, 메스나-프로파몰-아미오딜-클로피라미드-프로파몰-아미오딜-클로피라미드, 메스나-프로파몰-아미오딜-클로피라미드-프로파몰-아미오딜-클로피라미드-프로파몰, 메스나

• 후

- 어떤 요인으로 인해 후비루가 발생할 수 있는지 알고 싶어요?

QA 이후 결과 넣기

후비루는 다양한 요인에 의해 발생할 수 있습니다.

후비루의 원인은 알레르기비염, 비알레르기비염, 부비동염, 인후두 위산역류증, 연하장애, 피임약 사용, 호르몬 변화 등 다양한 요인에 의해 발생할 수 있습니다. 알레르기비염은 알레르기 원인물질에 노출될 때 발생하며, 비알레르기비염은 비알레르기 원인물질에 노출될 때 발생합니다. 부비동염은 부비동의 염증으로 인해 발생하며, 인후두 위산역류증은 위산 역류로 인해 발생합니다. 연하장애는 호흡기 질환으로 인해 발생하며, 피임약 사용은 호르몬 변화로 인해 발생할 수 있습니다. 후비루의 원인은 다양하며, 알레르기비염, 비알레르기비염, 부비동염, 인후두 위산역류증, 연하장애, 피임약 사용, 호르몬 변화 등 다양한 요인에 의해 발생할 수 있습니다.

Evaluation

Table 1: ASR (Key Performance Metrics by ASR Model)

Model	Latency (s)	RPS	RTF	Model size (GB)	CER (%)
voxtral 3B 4bit	1.21	0.83	5.86	2.18	19.14
whisper large FP16	0.75	1.32	7.74	2.87	8.66
ours	1.55	0.65	3.95	2.25	9.09

• ASR

- Latency : 오디오 샘플 1개를 처리하는 데 걸리는 평균 시간(초)
- RPS : 1초당 처리할 수 있는 오디오 샘플의 개수
- RTF : 실제 오디오 길이 대비 처리 속도
- CER : 변환된 텍스트가 정답과 문자 단위 오류율

Table 2: SQA(Speech Question Answering) GPT-OSS Model evaluation

Evaluation Model	Average Score (out of 5 points)
Voxtral 3B 4-bit (base)	2.07
Voxtral 3B 4-bit + RAG	2.91
Voxtral 3B 4-bit + LoRA (Ours)	4.01

• SQA

• LLM-as-a-Judge (GPT-OSS)

- ▶ 5점 (매우 우수): 질문의 핵심 의도를 완벽하게 이해하고, 매우 정확하고 상세하며 직접적인 답변을 제공함.
- ▶ 4점 (우수): 질문에 직접적으로 답변하며, 대부분의 내용이 정확하고 관련성이 높음. 약간의 아쉬움이 있을 수 있음.
- ▶ 3점 (보통): 질문의 일부에만 답변하거나, 답변의 관련성이 다소 떨어짐.
- ▶ 2점 (미흡): 질문의 핵심을 파악하지 못하고, 관련 없는 정보를 제공하거나 답변에 오류가 많음.
- ▶ 1점 (매우 미흡): 질문과 전혀 관련 없는 답변을 하거나, 완전히 잘못된 정보를 제공함.

Conclusion

- 단계적 Finetuning 전략의 효과
 - 텍스트 지식 학습이 ASR 성능 향상에도 기여함을 검증.
 - ASR 데이터셋에 없던 의료 용어까지 인식 가능.
 - Stage 1 (LLM LoRA, 텍스트 QA): 의료 지식 학습
 - Stage 2 (Audio Adapter + LLM LoRA, ASR+SQA): 음성 인식 + QA 적응
- 의료 발화 → 읽기 쉬운 문장으로 보정(추임새 제거·표현 정제)
- 실용성
 - 4-bit 양자화 + LoRA로 경량 실행 가능
 - 의료 현장의 프라이버시·지연 제약에 적합
- 향후 환자 맞춤형 설명·치료 과정 안내 등 확장 가능

Future Works

- 단계별 추론을 통한 성능 극대화
 - CoT(Chain-of-Thought) 프롬프팅을 도입하여 과업을 두 단계로 명확히 분리
 - 1단계: 음성 인식 오류를 먼저 교정
 - 2단계: 교정된 텍스트를 기반으로 최종 답변을 생성하는 파이프라인을 구축하여 답변의 정확성과 논리적 흐름을 개선
- 모델 정렬(Alignment)을 통한 최적화
 - RLHF, DPO와 같은 최신 모델 정렬 기법을 적용하여, 모델의 답변이 사용자의 의도에 더 잘 부합하고 유용하도록 최적화
 - 이 과정을 통해 불필요한 연산을 줄이고 모델의 응답 속도를 개선하는 연산 최적화도 함께 진행
- Streaming Inference
 - 실시간 발화 처리 및 부분 출력 지원
 - 의료 현장 대화에 적합한 실시간 응답성 확보

Reference

- [1] J. Jia et al., "Efficient Streaming LLM for Speech Recognition," *Meta AI*, 2024.
- [2] C. Xiao, Z. Hou, D. Garcia-Romero, and K. Han, "Contextual ASR with Retrieval Augmented Large Language Model," in *Proc. ICASSP*, 2025.
- [3] S. Li et al., "LA-RAG: Enhancing LLM-based ASR Accuracy with Retrieval-Augmented Generation," in *Proc. ICASSP*, 2025.
- [4] J. Peng et al., "A Survey on Speech Large Language Models for Understanding," *arXiv preprint*, 2024.
- [5] A. Gupta et al., "Improving Domain-Specific ASR with LLM-Generated Contextual Descriptions," *arXiv preprint*, 2024.
- [6] "WavRAG: Audio-Integrated Retrieval Augmented Generation for Spoken Dialogue Models," in *Proc. ACL*, 2025.
- [7] A. H. Liu et al., "Voxtral," *Mistral AI*, 2024.

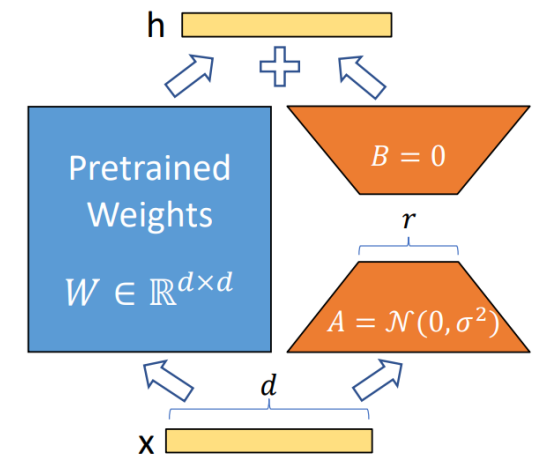
Additional tasks

- whisper-streaming 성능 평가
 - 의료 test set 활용
- 멀티턴 audio 데이터셋 normalize
 - Whisper large v3 모델 활용, CER 결과 확인 후 필터링
- 의료 text dataset 정리
 - Kenlm, Rnnlm 학습을 위한 데이터 병합 및 Normalize
 - 비대면 진료를 위한 의료진 및 환자 음성, 초거대 AI 헬스케어 질의응답 데이터, ASAN AMC (1M)
- kenlm 학습
- Contextual biasing research

Appendix

• QLoRA(4bit)

- 거대한 원본 모델은 4비트로 압축해서 '얼려두고(frozen)', 아주 작은 크기의 '어댑터(adapter)'만 추가해서 그 부분만 학습시키자!
- NF4(NormalFloat 4-bit)
 - ▶ LLM 가중치의 분포는 정규분포를 따름.
 - ▶ 가중치가 밀집된 0 주변에는 더 많은 지점을 촘촘하게 할당하고, 가중치가 드문 양쪽 끝 구간에는 지점을 드물게 할당.
- 4-bit NF4 Quantization
 - ▶ Voxtral-Mini-3B 전체를 이전과 동일하게 **4-bit NF4** 방식으로 양자화하여 VRAM에 로드
- LoRA (Low-Rank Adaptation)
 - ▶ 기존 Layer에 병렬로 Adapter를 추가
 - ▶ 두 개의 작은 행렬(A와 B)로 구성된 Adapter로 학습 파라미터 개수를 줄임
- Mixed Precision Training
 - ▶ '저장은 4비트, 계산은 16비트' 전략
 - ▶ NF4 codebook을 통해 16bit 대표값을 저장, 이들의 조합으로 가중치를 구성



Appendix

- Finetuning에서 QLoRA(4bit)를 사용하는 이유

- 고려사항

- Weights (Model Parameters)
- Optimizer
- Gradient
- Activations & Overhead

- 1. 메모리 요구량 (Voxtral mini 3B 모델 기준, 4,676,271,104) 최소 계산치

	Model Weights	Gradients	Optimizer	Activation	Sum
Full finetuning	9.4 GB	9.4 GB	37.6 GB	5GB	60 GB
LoRA (fp16)	9.43 GB	30 MB	120 MB	5GB	15 GB
QLoRA (4bit)	2.38 GB	30 MB	120 MB	5GB	8GB

Appendix

• Voxtral Mini 3B Streaming 가능성

1. 오디오 Chunk 단위 입력

- ▶ pyaudio/sounddevice로 0.5–1.0s 청크 수집 → processor.apply_transcription_request에 즉시 투입

2. 실시간 generate() + 부분 문장 디코딩

- ▶ 청크마다 max_new_tokens 제한으로 빠른 부분 출력(Partial) 갱신

3. 연속성 확보(문맥 유지)

- ▶ 직전 **b**개 청크만 텍스트 문맥에 포함(윈도우드 컨텍스트)
- ▶ 가능 시 KV-cache/hidden state 재사용으로 지연 최소화

• 참고(근거)

- [1] J. Jia et al., *Efficient Streaming LLM for Speech Recognition (SpeechLLM-XL)*, Meta AI, 2024.
 - ▶ 오디오를 청크로 나누어 EOS(청크 종료) 기반 순차 디코딩
 - ▶ 제한된 어텐션 윈도우(b 청크)로 계산량을 선형 수준으로 절감
 - ▶ 정렬 정보는 CTC 강제 정렬로 추정 → 실사용에 준하는 정렬 품질 확인