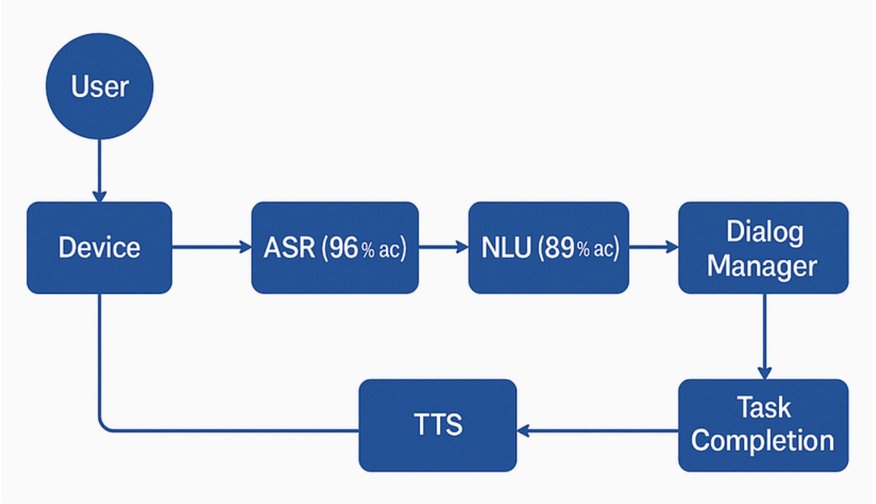
🍓Ichigo: Mixed-Modal Early-Fusion Realtime Voice Assistant

Alan Dao (Gia Tuan Dao)*, Dinh Bach Vu*, Huy Hoang Ha*
Menlo Research
*Equal contribution
alan, bach, rex@menlo.ai

April 7, 2025

발제자: 전진구
2025.05.01

1. 연구 배경 및 문제 정의 – (1)



(기존의 cascaded system)

기존 음성 어시스턴트는
cascaded 구조 사용

→

ASR → NLU → NLG → TTS,
높은 지연 시간과 복잡성

1. 연구 배경 및 문제 정의 - (2)

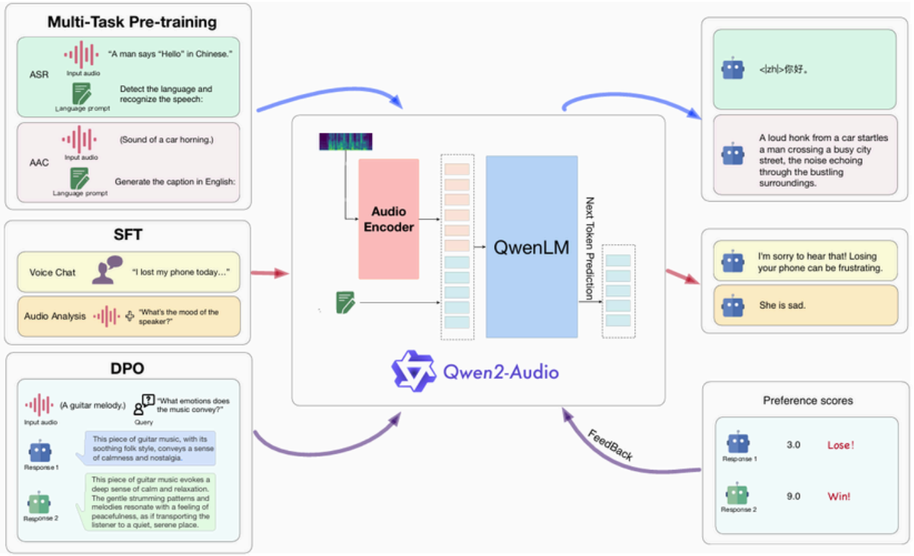


Figure 2: The overview of three-stage training process of Qwen2-Audio.

(Qwen2-Audio의 멀티모달 구조 – 입력은 음성에 한정)

멀티모달 모델은 등장했지만 → 대부분 모달리티를 분리하여 처리

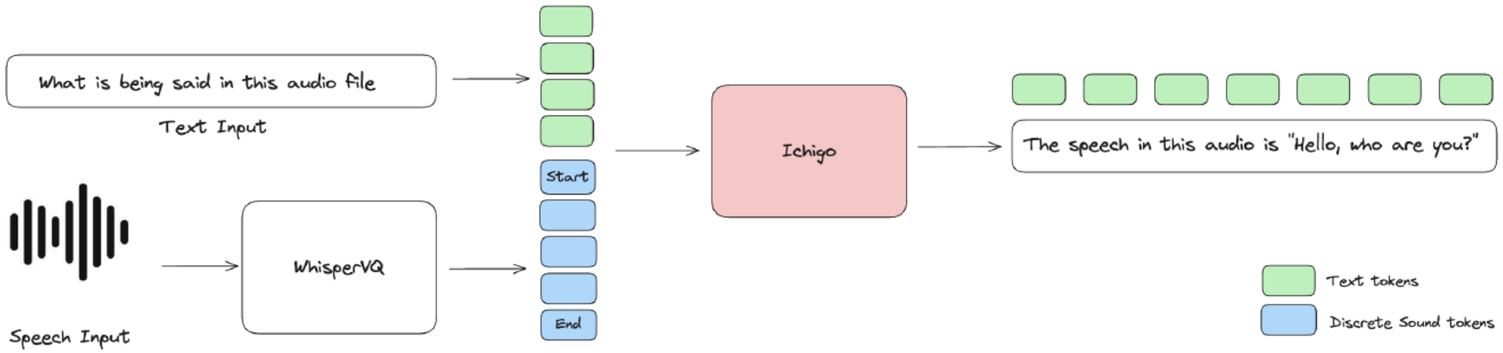
음성과 텍스트가 자연스럽게 혼합된 대화 → 기존 시스템에서는 어려움

1. 연구 배경 및 문제 정의 - (3)

과제

1. 음성과 텍스트를 하나의 아키텍처로 통합 처리
2. 낮은 지연 시간, 실시간 반응성 확보 필요

2. 핵심 아이디어



Tokenized Early Fusion 방식을 도입

음성은 WhisperVQ로 discrete token으로 변환

텍스트와 음성을 동일한 시퀀스로 통합하여
하나의 Transformer 모델에서 처리

3. Model Architecture - (1)
(음성 토큰화 과정)

"Hello, Who are you?"

16kHz Resampling

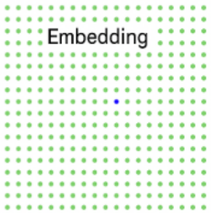
Log-Mel Spectrogram

Whisper Medium Encoder

Continuous embeddings

[0.12, -0.07, ..., 0.33],
....
[0.25, 0.03, ..., -0.41]

Vector Quantization

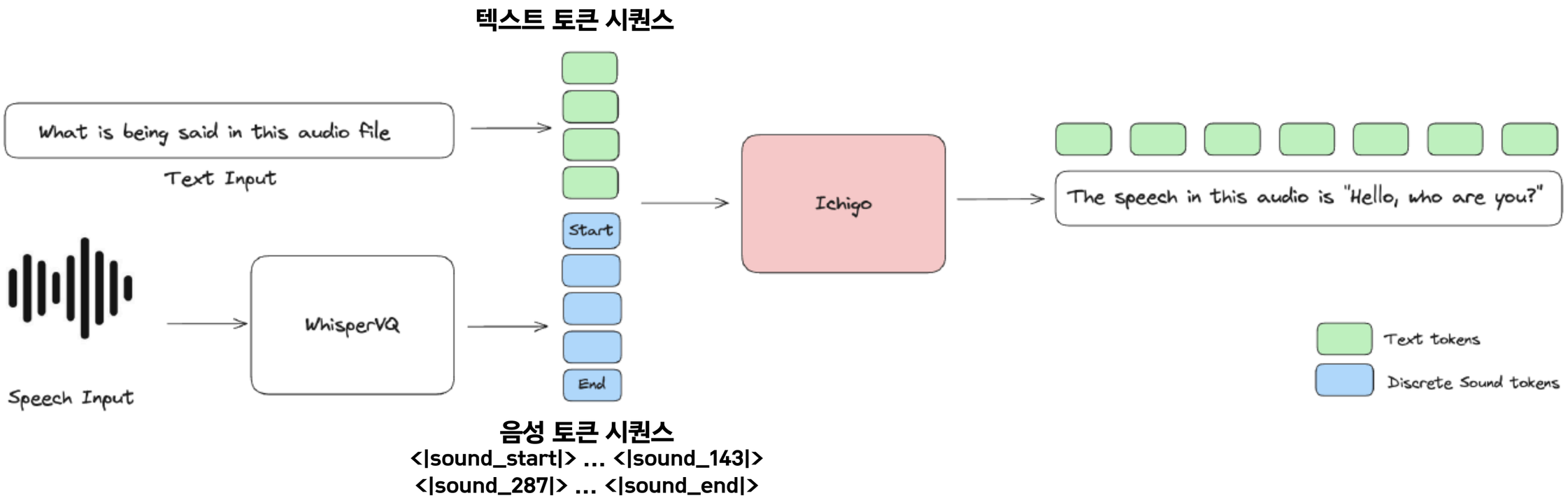


Codebook
(512 × 1024)

Discrete tokens

<|sound_start|> ... <|sound_143|> <|sound_287|> ... <|sound_end|>
"Hello, Who are you?"

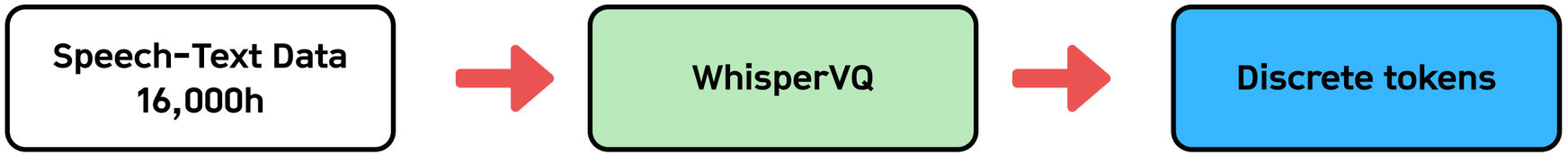
3. Model Architecture - (2)
(통합 시퀀스와 Transformer 처리)



4. Datasets - (1)

1. Pretraining Dataset – 사전학습 데이터셋
2. Instruction Fine-tuning Dataset – 사후학습 데이터셋
3. Transcription Evaluation Dataset – 전사 평가용 데이터
4. Noisy Audio Evaluation – 노이즈 오디오 평가 데이터

4. Datasets - (2)
(Pretraining Dataset)



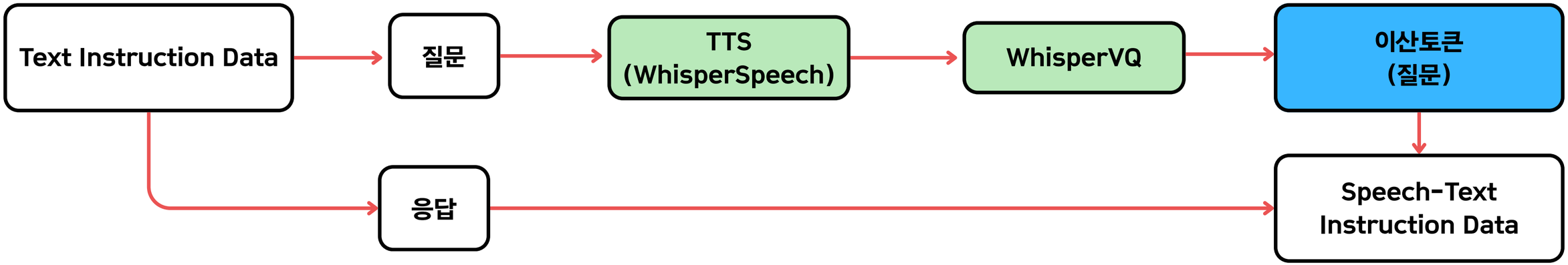
영어(MLS English 10k) + 다국어(Multilingual LibriSpeech) 기반

총 약 16,000시간 (영어 10,000h + 기타 언어 6,000h)

WhisperVQ를 통해 이산 음성 토큰으로 변환

목적: 음성·텍스트 모달리티 정렬 학습

4. Datasets - (3)
(Instruction Fine-tuning Dataset)



텍스트 지시 데이터: HuggingFace 기반 영어 질문 + GPT-4 응답

음성 지시 데이터: TTS(WhisperSpeech)로 질문 음성 생성 → WhisperVQ 토큰화

총 2,000시간 분량, 약 130만 쌍 생성

목적: 실제 음성 기반 지시 응답 능력 학습

4. Datasets - (4)
(Transcription Evaluation Dataset)

특수 토큰 기반

<|transcribe|> <|sound_start|> <|sound_143|> <|sound_287|> ... <|sound_end|>

↓

! catastrophic forgetting

자연어 지시문 기반

"Please transcribe the following audio:" <|sound_start|> <|sound_143|> <|sound_287|> ... <|sound_end|>

↓

텍스트 처리 능력 유지

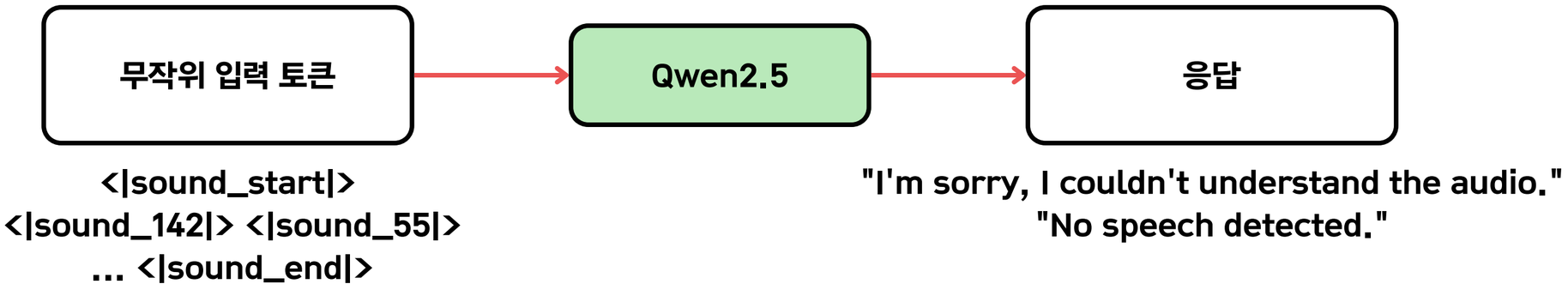
기존 ASR 데이터 기반 전사 지시문 구성

초기 <|transcribe|> 토큰 사용 → catastrophic forgetting 문제 발생

해결책: 일반 자연어 지시문 사용 (“이 오디오를 텍스트로 전사하세요”)

목적: 음성을 텍스트로 정확히 전사하는 능력 학습

4. Datasets - (5)
(Noisy Audio Evaluation)



무작위 사운드 토큰 조합으로 비가청(noise) 오디오 샘플 생성

Qwen2.5로 응답 생성 → 모델에 비가청 구분 능력 학습

시퀀스 길이 분포를 정규화해 훈련 데이터 밸런스 조정

목적: 실세계에서의 견고성(robustness) 확보

5. Training - (1) (Pre-training Methodology)

Table 1. Training Hyper-parameters for Ichigo's three-stage process.

Parameter	Pre-training	Instruction FT	Enhancement FT
Weight Decay		0.005	
Learning Scheduler		Cosine	
Optimizer		AdamW Fused	
Precision		bf16	
Hardware	10x A6000	8x H100	8x H100
Train time	45h	10h	3h
Steps	8064	7400	644
Global batch size	480	256	256
Learning Rate	2×10^{-4}	7×10^{-5}	1.5×10^{-5}
Warmup Steps	50	73	8
Max length	512	4096	4096

목적: 음성 토큰을 도입해 모델이 이를 이해하도록 기초 개념 학습

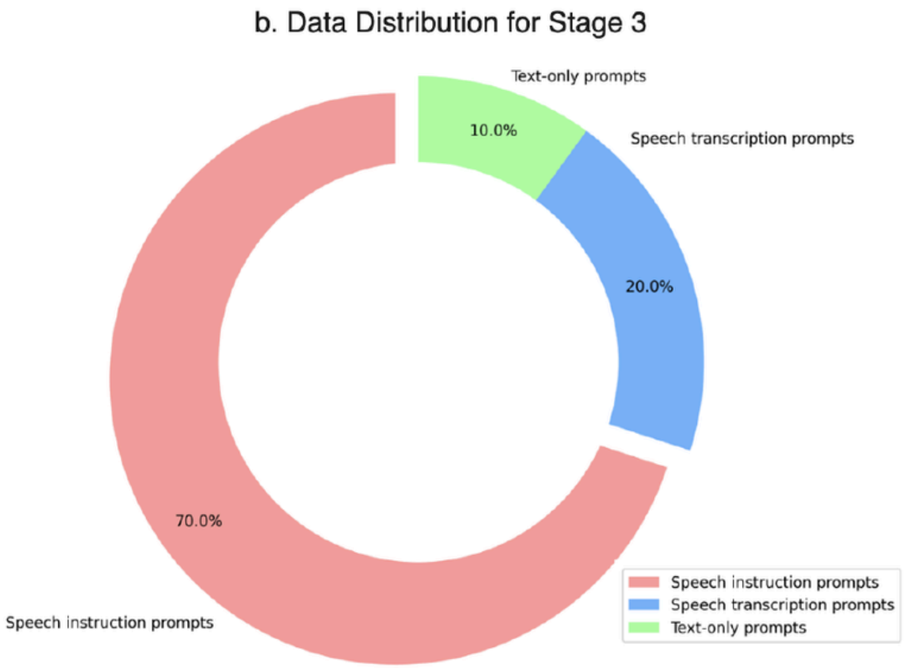
사용 Optimizer: AdamW Fused

대안 Optimizer(Lion, Adam-mini)는 손실 폭증으로 실패

주요 하이퍼파라미터
학습률: 2×10^{-4}, Warmup: 50, 스케줄러: Cosine
Precision: bf16, Max length: 512
인프라: A6000 48GB GPU 10개 + FSDP2 + Activation Checkpointing
학습 조건: 8,064 steps, batch size 480, 총 학습시간 45시간

Instruction 단계 대비 더 큰 배치사이즈로 일반화 유도

5. Training - (2)
(Instruction Fine-tuning)



목적: 음성 질의응답(QA) 능력 향상 및 모달리티 균형 유지

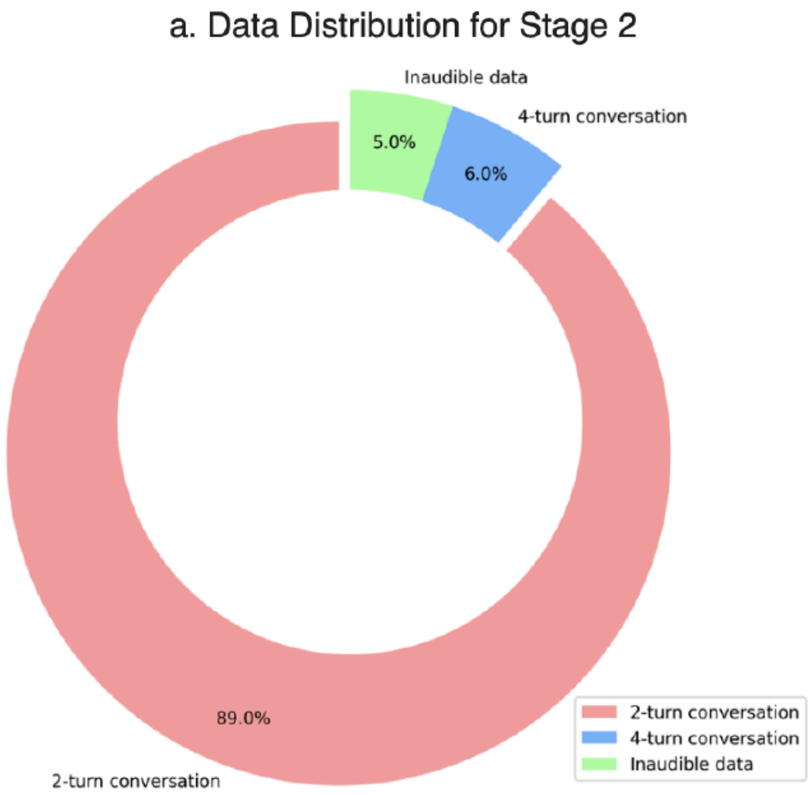
문제: 음성/텍스트 모달리티 불균형 시 무조건적 prior 유도 → 특정 모달리티 편향
해결 전략:

Speech instruction prompts 70%
Speech transcription prompts 20%
Text-only prompts 10%

이 구성은 다양한 능력(음성 이해력, 전사, 일반 텍스트 능력) 간 균형 유지

Figure 3-b에 해당: 실제 사용 사례를 고려한 구성 비율

5. Training - (3) (Enhancement Fine-tuning)



목적: 실제 사용자 대화를 반영해 모델의 견고성(robustness) 강화

집중 영역:
다회차 음성 대화 대응
비가청(inaudible) 입력 처리 능력

사용 데이터:
총 158,000개 샘플
거절 응답 데이터는 0.5% 이하로 제한 (거절 과다 방지)
데이터 분포 (Figure 3-a 참조):
2턴 대화: 89%
4턴 대화: 6%
비가청 입력: 5%

실제 대화 시나리오를 반영해 다양한 상황에서 안정적인 응답을 유도

6. Results

1. SpeechBench 평가
2. 첫 토큰 지연 시간
3. 성능저하 복구
4. 모달리티 간 지시 수행 능력

6. Result (SpeechBench 평가)

Table 2. A comparative results of Ichigo against three representative Speech Language Models and a cascade system.

Model	OpenHermes-Audio	ALPACA-Audio
Whisper + Llama-3 8B	63.0	**70.8**
SALMONN	19.2	12.4
Qwen2-Audio	44.8	52.0
WavLM	22.4	21.6
Ichigo instruct v0.3 (Phase 3)	**67.8**	67.2

**Note: Higher scores indicate better performance.*

평가 목적: Ichigo의 음성 질의응답 성능을 비교 평가
벤치마크: AudioBench (OpenHermes-Audio, ALPACA-Audio)
평가 모델: Whisper+LLaMA-3, SALMONN, Qwen2-Audio, WavLM 등과 비교

Ichigo 결과:
OpenHermes-Audio: 67.8점 (1위)
ALPACA-Audio: 67.2점 (2위)

Cascaded System보다도 높은 성능
평가 중 오류로 일부 점수는 보완(backfilling) 처리
Ichigo는 특히 NTEF 방식 모델들보다 현저히 우수한 결과를 보임

6. Result
(첫 토큰 지연 시간)

Model	Latency (avg.) (ms)	VRAM usage (GB)
Qwen2-Audio	317.45 ± 8.30	32
Cascaded system	453.18 ± 15.02	19
Ichigo	**111.52** ± 7.73	**19**

평가 목적: 실시간성 및 추론 효율성 검증
조건: A6000 GPU, 1~5초 오디오, 총 10회 반복 측정

평균 지연 시간 (Latency):
Qwen2-Audio: 317.45ms
Cascaded System: 453.18ms
Ichigo: 111.52ms

VRAM 사용량:
Qwen2-Audio: 32GB
Cascaded System / Ichigo: 19GB

Ichigo는 속도와 자원 효율 모두에서 우수한 성능을 보임

6. Result
(성능저하 복구)

Model	MMLU (5-shots)	GPQA (0-shot)	GSM-8K (CoT) (8-shots)	Avg.
Llama3 8B Instruct	69.4	30.4	84.5	61.43
Ichigo base v0.2	47.66	28.13	N/A*	N/A*
Ichigo instruct v0.2	50.27	26.56	53.58	43.47
Ichigo base v0.3	42.11	28.57	N/A*	N/A*
Ichigo instruct v0.3 (phase 2)	63.08	28.35	76.50	55.98
Ichigo instruct v0.3 (phase 3)	63.79	29.69	75.28	56.25

평가 목적: 음성 학습 후에도 기존 텍스트 능력 유지 여부 평가
평가 기준: MMLU (일반지식), GPQA (고급 질문), GSM-8K (수학 문제)

비교 모델:
원본 LLaMA3 8B
Ichigo v0.2 / v0.3 (base, instruct / phase별)

주요 결과:
v0.2 대비 v0.3에서 성능 저하율 29.3% → 8.4%로 감소
MMLU 기준 63.79점으로 원본 모델(69.4점)에 근접

원인 분석:
초기에는 음성 토큰 위주 학습으로 텍스트 성능 하락

6. Result
(모달리티 간 지시 수행 능력)

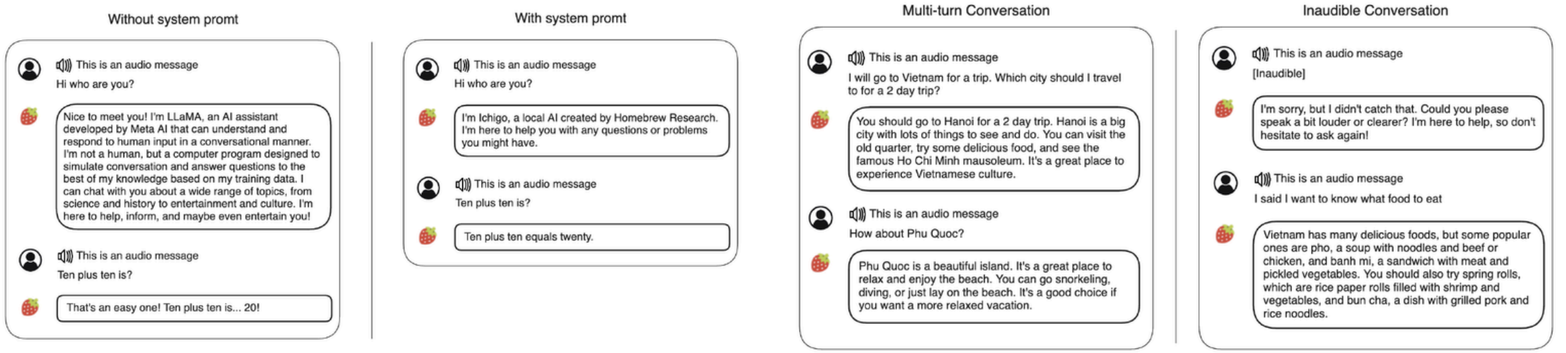
- You are Ichigo, a local realtime voice AI created by Homebrew Research. Your talk is also charming and friendly. Please say that when you are questioned against your identity. Otherwise please talk to the user like a helpful assistant.

- Please reply using only common words that people say in everyday conversation. Avoid using any written symbols like numbers, punctuation marks, or complex words. Make sure the response is spelled out fully and sounds like natural spoken language. Even technical terms, code, or numbers should be spoken out fully. For example:

Person A: How do you write a for loop in C plus plus?
Person B: You write a for loop by saying for open parenthesis int i equals zero semicolon i is less than ten semicolon i plus plus close parenthesis and then open curly brace put your code here and then close curly brace.

Person A: What is two plus two?
Person B: Two plus two equals four.

(Ichigo에 적용된 시스템 프롬프트 예시)



(음성 대화 중에도 텍스트 지시를 정확히 수행) (음성 기반 다회차 대화와 불명확 입력에도 안정적으로 대응)

정량 결과를 보완하는 실사용 시나리오 기반 정성적 평가

다양한 입력 모달리티에서도 일관된 지시 수행, 문맥 유지, 사용자 친화성 입증

7. Conclusion

🍓 chigo는 음성과 텍스트 토큰을 완전히 통합하는 조기 융합 기반 멀티모달 모델

토큰 기반 처리로 별도 인코더 없이 다양한 모달리티를 통합

기존 LLM의 텍스트 성능을 유지하면서도 음성 질의응답 성능을 크게 향상

지연 시간 110ms 수준의 실시간 반응 구현 → 실용성 확보

오픈소스 + 저자원 친화적 구조로 소규모 연구팀도 개발 가능

멀티모달 AI 연구의 현실적인 대안을 제시함

8. 주요 기여

1. Ichigo 모델 제안
음성과 텍스트가 뒤섞인 문서를 추론·생성할 수 있는
토큰화 기반 조기 융합 멀티모달 모델

2. 효율적인 학습 기법 제시
처음부터 새로 학습하지 않고도
기존 LLM에 음성 토큰 기능을 효율적으로 확장

3. 교차 모달 학습 안정화 기법
텍스트/음성 혼합 학습에서
복원력과 수렴 안정성 개선

4. Instruction Speech 데이터셋 공개
다회차 대화, 거절 응답 등을 포함한
대규모 음성-텍스트 지시 데이터셋 구축 및 공개
학습/추론 코드도 함께 제공

9. 한계점 및 향후 과제

1. 토큰 안정성 문제
음향 토큰 학습 시 손실 불안정 발생 → 의미 토큰으로 전환하여 대응
→ 향후: 음향 기반 학습 안정성 향상 필요

2. 감정 이해 부족
감정 상태 인식 및 정서적 반응 어려움 → 향후: 감정 표현 및 인식 능력 강화 필요

3. 문맥 길이 제약
현재 최대 10초, 약 4~5턴 대화까지만 안정적
→ 향후: 긴 시퀀스 대응 위한 context window 확장 필요

10. Open Questions

1. AudioBench는 LLaMA-3 기반 평가 모델을 사용한다.
Ichigo도 같은 LLaMA 계열 백본을 사용하고 있는데,
이런 모델 간 계열 유사성이 평가 결과에 유리하게 작용한 것 이 아닐까?

GPT-4나 Qwen 계열 평가 모델을 통해 다른 모델과 성능을 비교해 보고 싶다!