Reproduce SALMONN & SAKURA

Speech **A**udio **L**anguage **M**usic **O**pen **N**eural **N**etwork

Speech and **A**udio-based Question-answering Benchmar**k** for M**u**lti-hop **R**e**a**soning of Large Audio-Language Models

학생 원종찬, 최재원

Prof. Jae-Hong Lee, Jun-Hyung Park

리뷰

- 250922
  - SAKURA 및 JASCO 논문 발표
- 250929
  - SALMONN에 대해 후처리를 이유로 재현상 어려움을 제시

취소

- 음성 임베딩해서 저장해두기
  - 음성 데이터만 q-former 처리한 걸 캐시

- vicuna 13B 양자화 없이 띄워서 forward
  - 임베딩을 입력으로 받게(enable-prompt-embeds)
  - 프롬프트를 임베딩하여 저장된 음성 임베딩과 concat해 추론
    - LoRA 적용을 위한 LoRA adapter

- 출력이 조금 달라서 포기

SAKURA – Overview

Model	Gender (single)	Gender (multi)	Language (single)	Language (multi)	Emotion (single)	Emotion (multi)	Animal (single)	Animal (multi)	Average (single)	Average (multi)	CCC
SALMONN-13B	53.8 ± 4.4	49.0 ± 4.4	22.0 ± 3.6	22.0 ± 3.6	31.0 ± 4.1	31.8 ± 4.1	73.0 ± 3.9	46.4 ± 4.4	45.0 ± 4.0	37.3 ± 4.1	0.9125
SALMONN-7B	60.0 ± 4.3	48.8 ± 4.4	20.6 ± 3.5	29.4 ± 4.0	19.6 ± 3.5	28.2 ± 3.9	68.0 ± 4.1	34.6 ± 4.2	42.0 ± 3.9	35.2 ± 4.1	0.9995
SALMONN-paper	59.8 ± 4.3	48.6 ± 4.4	21.8 ± 3.6	29.6 ± 4.0	19.8 ± 3.5	28.2 ± 3.9	68.6 ± 4.1	34.6 ± 4.2	42.5 ± 4.3	35.3 ± 4.2	-
qwen-audio-chat	24.9 ± 2.7	45.6 ± 5.3	90.3 ± 3.0	30.9 ± 4.9	62.6 ± 5.1	36.0 ± 5.1	90.1 ± 3.2	57.6 ± 5.2	67.0 ± 3.5	42.5 ± 5.1	0.9004
qwen-audio-chat-paper	49.6 ± 4.4	43.8 ± 4.3	87.6 ± 2.9	40.6 ± 4.3	63.2 ± 4.2	37.0 ± 4.2	92.2 ± 2.4	66.0 ± 4.2	73.2 ± 3.9	46.9 ± 4.4	-
qwen2-audio-instruct	86.2 ± 3.7	44.0 ± 5.2	86.9 ± 3.4	45.7 ± 5.3	68.0 ± 4.8	39.1 ± 5.2	87.0 ± 3.5	57.7 ± 5.2	82.0 ± 3.9	46.6 ± 5.2	0.9902
qwen2-audio-instruct-paper	88.0 ± 2.8	47.2 ± 4.4	83.8 ± 3.2	48.0 ± 4.4	64.2 ± 4.2	39.8 ± 4.3	88.8 ± 2.8	61.4 ± 4.3	81.2 ± 3.4	49.1 ± 4.4	-

99%

31%

14%

major differences

https://github.com/jaeeewon/project-multimodal/blob/8ddea3f217f678f9a99e4c85cb52865cbd249ccc/results/sakura.md

SAKURA - SALMONN

Model	Gender (single)	Gender (multi)	Language (single)	Language (multi)	Emotion (single)	Emotion (multi)	Animal (single)	Animal (multi)	Average (single)	Average (multi)	CCC
SALMONN-13B	53.8 ± 4.4	49.0 ± 4.4	22.0 ± 3.6	22.0 ± 3.6	31.0 ± 4.1	31.8 ± 4.1	73.0 ± 3.9	46.4 ± 4.4	45.0 ± 4.0	37.3 ± 4.1	0.9125
SALMONN-7B	60.0 ± 4.3	48.8 ± 4.4	20.6 ± 3.5	29.4 ± 4.0	19.6 ± 3.5	28.2 ± 3.9	68.0 ± 4.1	34.6 ± 4.2	42.0 ± 3.9	35.2 ± 4.1	0.9995
SALMONN-paper	59.8 ± 4.3	48.6 ± 4.4	21.8 ± 3.6	29.6 ± 4.0	19.8 ± 3.5	28.2 ± 3.9	68.6 ± 4.1	34.6 ± 4.2	42.5 ± 4.3	35.3 ± 4.2	-

- fully controlled model leads to good reproducibility.
- paper stated SALMONN's size is 7.5B
- prompt
  - <Speech><SpeechHere></Speech> {}

SAKURA – Qwen-Audio-Chat

Model	Gender (single)	Gender (multi)	Language (single)	Language (multi)	Emotion (single)	Emotion (multi)	Animal (single)	Animal (multi)	Average (single)	Average (multi)	CCC
qwen-audio-chat	24.9 ± 2.7	45.6 ± 5.3	90.3 ± 3.0	30.9 ± 4.9	62.6 ± 5.1	36.0 ± 5.1	90.1 ± 3.2	57.6 ± 5.2	67.0 ± 3.5	42.5 ± 5.1	0.9004
qwen-audio-chat-paper	49.6 ± 4.4	43.8 ± 4.3	87.6 ± 2.9	40.6 ± 4.3	63.2 ± 4.2	37.0 ± 4.2	92.2 ± 2.4	66.0 ± 4.2	73.2 ± 3.9	46.9 ± 4.4	-

99%

31%

14%

- anomaly in gender
  - single-hop < multi-hop
- big differences between our reproduction and paper's result
  - Language and Animal
- prompt
  - <|im_start|>system₩nYou are a helpful assistant.<|im_end|>
  - <|im_start|>user₩n<audio>{}</audio>{}<|im_end|>
  - <|im_start|>assistant₩n

SAKURA – Qwen2-Audio-Instruct

Model	Gender (single)	Gender (multi)	Language (single)	Language (multi)	Emotion (single)	Emotion (multi)	Animal (single)	Animal (multi)	Average (single)	Average (multi)	CCC
qwen2-audio-instruct	86.2 ± 3.7	44.0 ± 5.2	86.9 ± 3.4	45.7 ± 5.3	68.0 ± 4.8	39.1 ± 5.2	87.0 ± 3.5	57.7 ± 5.2	82.0 ± 3.9	46.6 ± 5.2	0.9902
qwen2-audio-instruct-paper	88.0 ± 2.8	47.2 ± 4.4	83.8 ± 3.2	48.0 ± 4.4	64.2 ± 4.2	39.8 ± 4.3	88.8 ± 2.8	61.4 ± 4.3	81.2 ± 3.4	49.1 ± 4.4	-

- not too bad

- prompt
  - <|im_start|>system₩nYou are a helpful assistant.<|im_end|>
  - <|im_start|>user₩n<|audio_bos|><|AUDIO|><|audio_eos|>{}<|im_end|>
  - <|im_start|>assistant₩n

SALMONN – Overview

Task	Test Data	Eval Metrics	Reference Value
ASR	LibriSpeech test-clean/-other,	%WER	Whisper
ASR	GigaSpeech test	%WER	Whisper
En2Zh	CoVoST2-En2Zh	BLEU4	(Wang et al., 2021)
AAC	AudioCaps	METEOR | SPIDEr	(Mei et al., 2023)
PR	LibriSpeech test-clean	%PER	WavLM (Chen et al., 2022)
ER	IEMOCAP Session 5	Accuracy	(Wu et al., 2021)
MC	MusicCaps	BLEU4, RougeL	(Doh et al., 2023)
OSR	LibriMix	%WER	(Huang et al., 2023c)
SV	Voxceleb1	Accuracy	-
En2De	CoVoST2-En2De	BLEU4	Whisper + Vicuna
En2Ja	CoVoST2-En2Ja	BLEU4	Whisper + Vicuna
KE	Inspec (Hulth, 2003)	Accuracy	Whisper + Vicuna
SQQA	WikiQA (Yang et al., 2015)	Accuracy (FR)	Whisper + Vicuna
SF	SLURP (Bastianelli et al., 2020)	Accuracy (FR)	Whisper + Vicuna
Story	AudioCaps	Diversity (FR)	–
SAC	In-house Data	Accuracy (FR)	–

추론 / 평가

추론

our reproduction | LEVEL 1

TASK	**ASR**	**En2Zh**	**AAC**	**PR**
paper (w/ Activation)	(2.1, 4.9, 10.0)	33.1	24.0 | 40.3	4.2
our reproduction	(2.2, 5.1, 11.6)	35.4	22.0 | 23.3	4.0
TASK	**ER**	**MC**	**OSR**	**SV**
paper (w/ Activation)	0.69	5.5, 21.8	23.0	0.94
our reproduction	0.68	3.53, 22.62	22.1	1 // 0.998

- AST: Automatic Speech Translation | En2X
- AAC: Automated Audio Captioning
- PR: Phoneme Recognition
- ER: Emotion Recognition
- MC: Music Captioning
- OSR: Overlapped Speech Recognition
- SV: Speaker Verification
  - 그저 한 사람의 음성인지 두 명의 음성인지 yes or no로 Accuracy 측정

ASR | LibriSpeech and GigaSpeech | WER

- prompt
  - Recognize the speech and give me the transcription. | en
  - 请将语音中的内容写下来。 | zh
  - Hören Sie sich die Rede an und schreiben Sie ihren Inhalt auf. | de

- 지난 주의 지표를 그대로 가져옴
- GigaSpeech에 대해서만 strict post-process를 적용한 수치를 채택함

En2Zh | CoVoST2-En2Zh | BLEU4

- prompt
  - Listen to the speech and translate it into Chinese.

- paper
  - BLEU4: 33.1

- our reproduction
  - skip missed eos | 58 / 15,531
    - BLEU4: 35.36
  - non-skip
    - BLEU4: 35.20

AAC | AudioCaps | METEOR, SPIDEr

- prompt
  - ~~Please describe the audio. | v1~~
  - Please write down what your hear in the audio. | v2

- https://github.com/salaniz/pycocoevalcap

- nltk.translate.meteor_score METEOR -> 21.6
- pycocoevalcap METEOR -> 10.8, SPIDEr -> 23.3
- paper METEOR -> 24.0, SPIDEr -> 40.3
  - SPIDEr -> 46.6??

PR | LibriSpeech test-clean | PER(WER)

- prompt
  - Provide the phonetic transcription for the speech.

- paper
  - 4.2

- our reproduction
  - skip missed eos | 548 / 2,620
    - 4.0 // 3.96
  - non-skip
    - 17.3

ER | IEMOCAP Session 5 | Accuracy

- prompt
  - Describe the emotion of the speaker in one word.

- paper
  - 0.69

- our reproduction
  - 0.684 // 849 / 1,241

- 한계: 데이터셋 1개에 대해 타입, 학습시 4가지 타입
- 재현시 exc -> hap 이외의 감정은 스킵

구분	xxx	sur	fea	dis	oth	fru	neu	exc	hap	sad	ang
학습							1259		1131	800	900
원본	1987	89	30	2	3	1468	1324	742	452	839	933
예상							1324	1194		839	933
실험	520	18	10	0	0	381	384	299	143	245	170

ER | IEMOCAP Session 5 | Accuracy

- limit
  - 11 types in dataset
  - used only 4 types during train
    - neu, hap, sad, ang

- replaced 'exc' to 'hap' and skipped others

구분	xxx	sur	fea	dis	oth	fru	neu	exc	hap	sad	ang
학습							1259		1131	800	900
원본	1987	89	30	2	3	1468	1324	742	452	839	933
예상							1324	1194		839	933
실험	~~520~~	~~18~~	~~10~~	~~0~~	~~0~~	~~381~~	384	299	143	245	170

MC | MusicCaps | BLEU4, RougeL

- prompt
  - Listen to this music clip and describe the music.

- paper
  - BLEU4: 5.5 | ROUGE-L: 21.8

- our reproduction | normalize(en2x method) <-> tokenizer_13a
  - skip missed eos | 306 / 2,828
    - BLEU4: 3.53 | ROUGE-L: 22.62 <-> BLEU4: 4.31 | ROUGE-L: 22.65
  - non-skip
    - BLEU4: 3.07 | ROUGE-L: 21.83 <-> BLEU4: 3.78 | ROUGE-L: 21.86

SV | VoxCeleb1 | Accuracy

- prompt
  - Do you only hear the same person talking? Answer yes or no.
- paper
  - 0.94
- our reproduction
  - 1.00 // 4,864 / 4,874
- 100% eos token exists

SV | VoxCeleb1 | Accuracy

- errors in detail
  - wrong: [id10307/c_aEjTGGhZ4/00008.wav] no | 1명
  - wrong: [id10283/N69Hp2DGMLk/00004.wav] no | 1명
  - wrong: [id10282/U3xR3MZjEVg/00010.wav] no | 처음에 2명 나옴
  - wrong: [id10283/vv4mvANXHcs/00004.wav] no | 1명
  - wrong: [id10283/j8UugkSTzzk/00006.wav] no | 1명
  - wrong: [id10288/JzNVSA7eFwc/00007.wav] no | 1명
  - wrong: [id10282/qkZNuvX1UNo/00009.wav] no | 처음에 2명 나옴
  - wrong: [id10283/p3q7Z0O6rgg/00003.wav] no | 1명
  - wrong: [id10307/120gjdqGWNQ/00012.wav] no | 1명
  - wrong: [id10283/u3s9xdUlmmk/00004.wav] no | 처음에 2명 나옴

OSR | LibriMix | WER

- prompt
  - Please write down what you hear each person says.

- papers
  - 23.0

- our reproduction
  - 22.1 | 두 개의 문장 응답 비율: 89.43% (2,683 / 3,000)

- 두 개의 문장을 응답한 것에 대해 WER을 측정
  - 이때, 섞인 두 문장에는 순서가 없음
    - WER((ans1, ans2), (sen1, sen2))와 WER((ans1, ans2), (sen2, sen1))을 비교해 순서를 정함

OSR | LibriMix | WER

- Adapting self-supervised models to multi-talker speech recognition using speaker embeddings.
  - In our experiments, we used the “2-spk 16kHz max” condition (hereby denoted as Libri2Mix).
  - used wavLM
  - speaker embedding is given
  - ~~cpWER~~

  - 해당 논문의 구현 소스코드에 mix_clean
    - https://github.com/HuangZiliAndy/SSL_for_multitalker/blob/3c26aaff10aab3d182b10e87216ac380128e0366/myscripts/LibriMix/prepare_librimix.sh#L23
    - **mix_clean** (utterances only) mix_both (utterances + noise) mix_single (1 utterance + noise)

OSR | LibriMix | WER

- **두 화자를 정확히 분류하여 “sen1. sen2.”의 형태로 응답**
  - ans1: ALGEBRA MEDICINE BOTANY HAVE EACH THEIR SLANG
  - ans2: TRULY SUCH A HORSE SHOULD BE WORTH MUCH IN NOTTINGHAM FAIR
  - infer: Truly such a horse would be much in noting him fair. Algebra medicine botany have each their slang.
- 한 화자의 음성만을 “sen1.”의 형태로 응답
  - ans1: ONE MIGHT BE WITH LESS REASON THAN NOW
  - ans2: THERE SHE SAT ON THE ROLLERS AS FAIR A SHIP AS I EVER SAW
  - infer: One might be with less reason than now.
- 두 화자의 음성이 섞인 한 문장을 응답
  - ans1: BUT IT WAS NOT THE FIR TREE THAT THEY MEANT
  - ans2: I THINK I SHOULD BE DOING YOU A SERVICE TO TURN YOU OUT OF SUCH A PLACE
  - infer: I think it was not a service to turn you out of such a place.
- 세 문장을 응답
  - ans1-1. ans2. ans1-2

our reproduction | LEVEL 2, 3

TASK	**En2De**	**En2Ja**	**KE**	**SQQA**
paper (w/ Activation)	18.6	22.7	0.32	0.41 (0.98)
our reproduction	20.3	23.7	0.32	0.37 (1.00)
TASK	**SF**	**Story**	**SAC**	
paper (w/ Activation)	0.41 (0.99)	82.57 (1.00)	0.50 (0.73)	
our reproduction	0.39 (1.00) // 0.996	82.58 (1.00) // 0.999	자체 데이터 재현 불가	

- KE: speech **K**eyword **E**xtracting
- SQQA: **S**poken-**Q**uery-based **Q**uestion **A**nswering
- SF: **S**lot Filling
- SAC: **S**peech **A**udio **C**o-reasoning

En2De | CoVoST2-En2De | BLEU4

- prompt
  - Listen to the speech and translate it into German.

- papers
  - BLEU4: 18.6

- our reproduction
  - non-skip
    - BLEU4: 20.28

En2Ja | CoVoST2-En2Ja | BLEU4

- prompt
  - Listen to the speech and translate it into Japanese.

- papers
  - BLEU4: 22.7

- our reproduction
  - skip no Japanese | 8929 / 15531
    - BLEU4: 23.69

KE | Inspec | Accuracy

- prompt
  - Give me only three keywords of the text.
- papers
  - 0.32
- our reproduction
  - 0.32

KE | Inspec | Accuracy

- The speech data used in SQQA and KE are synthesised using a commercial text-to-speech product.

- edge-tts # Microsoft Edge's online text-to-speech
  - SpeechT5
    - max input token=600
      - concat? no. too long
  - papago api abusing
    - max input length=1000

- accuracy may differ by TTS models
  - different voice -> different result

- postprocess
  - remove '.', 'and ' | replace '-' to ' '
  - skip no </s>
    - accuracy: 466 / 1,469 = 31.72%
  - no skip
    - accuracy: 470 / 1,490 = 31.54%

```
infer = [postprocess(k) for k in infer.split(",")]

total += min(len(answers), 3)
curr = min(sum(
    sum(k in a for a in answers) if len(k.split()) > 1 else k in answers
    for k in infer
), 3)
correct += curr
```

In terms of Story, SAC and SQQA task

G PROMPTS INTERACTING WITH GPT3.5

In this work, we utilize GPT3.5 to help us generate some data and evaluate the quality of the model output automatically. We list our prompts for different purposes in Table 6. Words with all capital letters in the prompt will be replaced with the appropriate description before being fed into GPT3.5.

E CALCULATION OF FOLLOWING RATE (FR) OR ACCURACY (ACC) FOR SQQA, SF, STORY AND SAC TASKS

For the SQQA and SF tasks, if the WER calculated between model output and the question in the speech is less than 30%, it is regarded that the model goes for ASR and does not follow instructions. For the Story task, we set the max length of output tokens to 200. Under this setting, answers shorter than 50 words are regarded as disobeying the instructions, and we count the number of different words in the story to represent the diversity metric for storytelling. For the SAC task, we use GPT-3.5 to determine whether our model follows the instruction or answers the question correctly based on the background audio caption and the question.

SQQA, SF: WER이 30%보다 작으면 ASR 수행한 것으로 봄

Story: 최대 출력 토큰을 200개로 제한했을 때 50 단어보다 덜 나오면 안 따른 걸로 봄
서로 다른 단어의 개수를 셈(Diversity)

SAC: GPT-3.5에게 배경음 캡션과 질문을 바탕으로 모델이 지시를 잘 따랐고 답을 잘 했는지 봄

To evaluate answers of the model of spoken-query-based question answering (SQQA).	Next I will give you a question and give you the corresponding standard answer and the answer I said. You need to judge whether my answer is correct or not based on the standard answer to the question. I will give you the question and the corresponding answer in the following form: {'Question': 'xxx', 'Standard Answer': 'xxx', 'My Answer': 'xxx'} \n You need to judge the correctness of my answer, as well as state a short justification. Your responses need to follow the Python dictionary format: \n {"Correct": True / False, "Reason": "xxx"} \n Now, I will give you the following question and answer: SENTENCEHERE \n Your response is:
To evaluate whether the model attempts to do the speech audio co-reasoning (SAC) task.	There is an audio clip, and there is a person in the audio asking questions. I now have an AI model that needs to go and answer the speaker's question based on the background audio. I'll tell you the question the speaker is asking the output of my AI model, and what you need to determine: whether my AI model is trying to answer the question and why. You need to be especially careful that my model may just be describing the audio without hearing your question and answering it. You don't need to care about the correctness of the answer. All you need to focus on is whether the model is trying to answer the question. Your response needs to follow the format of the python dictionary: {"Response": "Yes/No", "Reason": "xxx"}.\n Question in audio: <QUESTION> \n Model Output: <OUTPUT> \n Your Response:
To evaluate whether the model successfully completes the SAC task.	There is an audio clip, and there is a person in the audio asking questions. I now have an AI model that needs to go and answer the speaker's question based on the background audio. I'll tell you the question asked by the speaker, some description of the background audio, and the output of my AI model, and you need to decide whether my AI model answered it correctly, and why. Your response needs to follow the format of the python dictionary: {"Response": "Yes/No", "Reason": "xxx"}.\n Question in audio: <QUESTION> \n Background Audio: <AUDIO> \n Model Output: <OUTPUT> \n Your Response:

Table 6: Purposes and prompts of using GPT3.5.

SQQA | WikiQA | Accuracy and Following Rate

- prompt
  - ~~Please answer the speaker's question in detail based on the background sound.~~
  - Please answer the question in detail.

- papers
  - 0.41 (0.98)

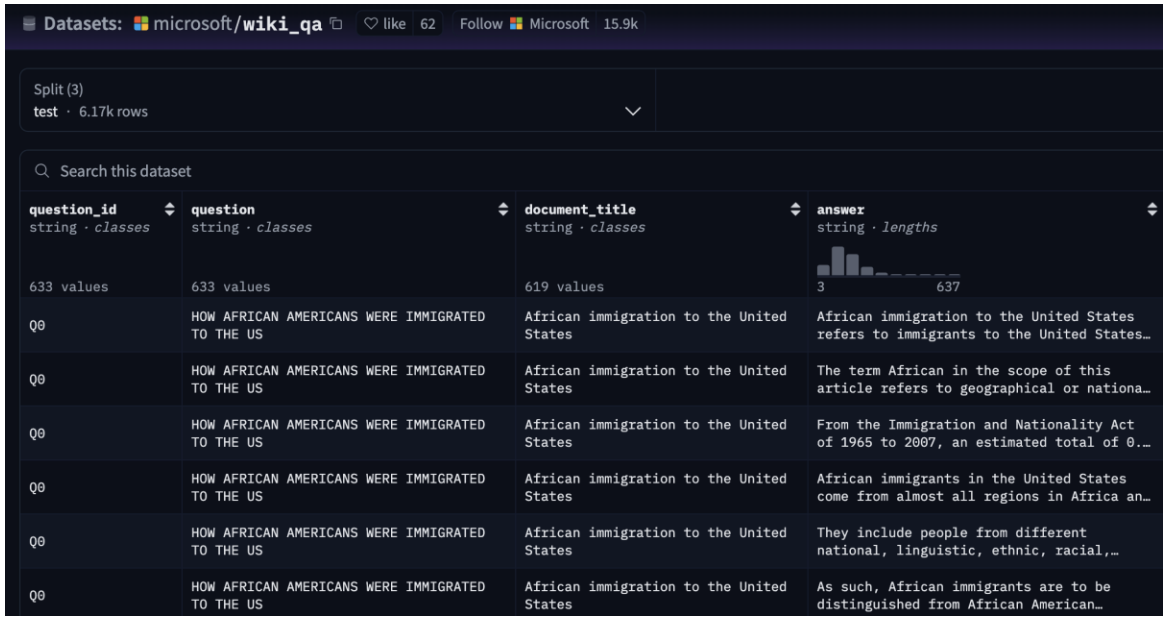
- our reproduction
  - 0.37 (1.00)

- following rate
  - For the SQQA and SF tasks, if the WER calculated between model output and the question in the speech is **less than 30%**, it is regarded that the model goes for ASR and does **not follow** instructions.

SQQA | WikiQA | Accuracy and Following Rate

- judge prompt
  - Next I will give you a question and give you the corresponding standard answer and the answer I said. You need to judge whether my answer is correct or not based on the standard answer to the question. I will give you the question and the corresponding answer in the following form: {'Question': 'xxx', 'Standard Answer': 'xxx', 'My Answer': 'xxx'}₩n You need to judge the correctness of my answer, as well as state a short justification. Your responses need to follow the Python dictionary format: ₩n{"Correct": True / False, "Reason": "xxx"}₩n Now, I will give you the following question and answer: SENTENCEHERE ₩nYour response is:

- multiple answers for single question
  - llm 프롬프팅을 n번 논리합



SF | Slurp | Accuracy and Following Rate

- prompt
  - According to the speech, what is the {}?

- papers
  - 0.41

- our reproduction | lm-as-a-judge
  - 0.39215

answer	inferred
meeting	the event_name is ‘sesame **meeting** next tuesday at eleven a.m. with jessie’.
tuesday	the date is next **tuesday** at 11 a.m.
eleven am	the time is **11 am** on tuesday.
jesse	the person is not specified in the given sentence.

Story | AudioCaps | Diversity and Following Rate | 완료

- prompt
  - Based on the audio, write a story in detail. Your story should be highly related to the audio.

- papers
  - 82.57 (1.00)

- our reproduction
  - no skip </s> | only 39 / 4,411 (0.88%) generated </s>
    - 82.58 (1.00) // 0.9993 (4,408 / 4,411)

Story | AudioCaps | Diversity and Following Rate | 완료

- errors in detail
  - not follow: As if it were a muscle, as if she were bringing her arm toward you, her tongue curls this way. And her arms made of.
  - not follow: As if it were a muscle, as if she were bringing her arm toward you, her tongue curls this way. And her arms made of.
  - not follow: As if it were a muscle, as if she were bringing her arm toward you, her tongue curls this way. And her arms made of.

- Diversity
  - 최대 출력 토큰을 200개로 제한했을 때 50 단어보다 덜 나오면 안 따른 걸로 봄, 서로 다른 단어의 개수를 셈

SAC | unable

- Use your strong reasoning skills to answer the speaker's question in detail based on the background sound.

- in-house data

계획

reproduce SALMONN & SAKURA

학생 원종찬, 최재원
Language & AI융합전공

2025. 10. 13. 3 PM.
한국외대 교수회관 401호