Reproduce SALMONN

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

학생 원종찬, 최재원

Prof. Jae-Hong Lee, Jun-Hyung Park

Environments

- RTX5090 * 4
- RTX4090 * 4

Environments

🌈 How to inference in CLI

1. Same as **How to train a model: 1-4**.
2. Download salmonn v1 to `ckpt`.
3. Running with `python3 cli_inference.py --cfg-path configs/decode_config.yaml` in A100-SXM-80GB.
   Now you can input `wav_path` and `prompt`. Enjoy yourself !

Environments

- major difference
  - 8bit quantization for Vicuna LLM

2025년 9월 24일 오전 3:02

quantization comparison

/home/jpong/Workspace/jaeeewon/repr_salmonn/salmonn/resource/audio_demo/duck.wav
Describe the following audio in a caption.

4bit | torch.float16, torch.bfloat16 | Ducks are quacking and a man is singing.</s>
8bit | torch.float16 | A woman sings while ducks quack in the background.</s>
org | OOM

/home/jpong/Workspace/jaeeewon/LibriSpeech/test-other/3080/5040/3080-5040-0000.flac
Can you transcribe the speech into a written format?

4bit | Would it would leave me and then i could believe i shall not always have occasion for it</s>
8bit | Would it would leave me and then i could believe i shall not always have occasion for it</s>
org | Would it would leave me and then i could believe i shall not always have occasion for it</s>

/home/jpong/Workspace/jaeeewon/LibriSpeech/test-other/3080/5040/3080-5040-0000.flac
Describe the following audio in a caption.

4bit | The audio is a female voice reciting a poem.</s>
8bit | The audio is a monologue of a person reflecting on the possibility of leaving something and not having the opportunity to do so again.</s>
org | The audio is a monologue of a person reflecting on the possibility of leaving something and not having the opportunity to do so again.</s>

/home/jpong/Workspace/jaeeewon/repr_salmonn/salmonn/resource/audio_demo/duck.wav
Can you transcribe the speech into a written format?

4bit | Yes, I can transcribe the speech into a written format.</s>
8bit | Yeah, you want to take your duck call and say.</s>
org | OOM

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

PostProcessing

- postprocessing not mentioned

PostProcessing – non-strict | applied


```
if skip and not infer.endswith("</s>"):
    curr_missed_eos += 1
    continue
```

PostProcessing – strict | not applied


```python
answer = task_data.get("sentence").strip()
infer = task_data.get("infer").strip()

if not (infer.endswith("<s>") or infer.endswith("</s>")):
    status = "no_eos"
elif infer.startswith("The transcription of the given speech is:"):
    status = "garbage_prefix_v1"
    infer = (
        infer[len("The transcription of the given speech is:") :]
        .strip()
        .split("\n")[0]
    )
elif infer.startswith("The transcription of the speech is:"):
    status = "garbage_prefix_v4"
    infer = (
        infer[len("The transcription of the speech is:") :].strip().split("\n")[0]
    )
elif "<Speech>" in infer and "</Speech>" in infer:
    # ~<Speech>the making of a loyal patriot</Speech>~일 경우, <Speech> </Speech> 안의 내용만 추출해야
    status = "speech_tagged"
    infer = infer.split("<Speech>", 1)[1].split("</Speech>", 1)[0].strip()
elif infer.startswith("I'm sorry"):
    status = "sorry_rejected"
elif ': "' in infer:
    # Here is the transcription of the speech: "the villages will save us in the end"</s>
    # Here is the transcription: "This snatcher had been an orphan for many years."
    status = "garbage_prefix_v2"
    infer = infer.split(': "', 1)[1].split('"', 1)[0].strip()
    # transcription이 두 번 나와서 앞에서도 자르고 뒤에서도 잘라야 함
elif 'is "' in infer:
    status = "garbage_prefix_v3"
    infer = infer.split('is "', 1)[1].rsplit('"', 1)[0].strip()
    # The speech is: "This snatcher had been an orphan for many years."\n\nHere is the transcription: "This snatcher had been an orphan for many years."
elif abs(len(answer) - len(infer)) > 10:
    status = "too_long"
elif infer.endswith("</s>"):
    status = "valid"
else:
    status = "invalid"
```

Overview – ASR w/o non-strict post-processing

- LibriSpeech test-clean | 2,620
  - 0.02 (2.24%)
- LibriSpeech test-other | 2,939
  - 0.05 (5.12%)
- GigaSpeech test | 25,619
  - 0.56 (56.39%)
- overall | 31,178
  - 0.45 (45.27%)

Method	**ASR↓**
w/o Activation	(2.1, 4.9, 9.1)
w/ Activation	(2.1, 4.9, 10.0)
Reference Value	(2.2, 5.1, 9.2)

Overview – ASR w/ non-strict post-processing

- LibriSpeech test-clean | 2,620
  - 0.02 (2.24%)
- LibriSpeech test-other | 2,939
  - 0.05 (5.12%)
- GigaSpeech test | 25,619
  - 0.12 (11.56%)
- overall | 31,178
  - 0.10 (9.89%)

Method	**ASR↓**
w/o Activation	(2.1, 4.9, 9.1)
w/ Activation	(2.1, 4.9, 10.0)
Reference Value	(2.2, 5.1, 9.2)

Overview – PR data

How are the phone sequences obtained #103

Closed

cantabile-kwok opened on Mar 23

Hi! SALMONN supports phoneme recognition, and the generated phones seem to be CMU dict phones with optional silence. Hence, I wonder how is these phones generated? Are they generated from forced alignment tools like MFA? Also, are the phone sequences for LibriSpeech dev and test set available online?

Thank you very much.

TCL606 on Mar 24 — Collaborator

Sorry, I made a mistake. LibriSpeech itself does not contain phoneme annotations. We use MFA to get the annotations.

TCL606 on Mar 24 — Collaborator

LS_devclean_pr.json
LS_testclean_pr.json

src: https://github.com/bytedance/SALMONN/issues/103

The CMU Pronouncing Dictionary

- Look up the pronunciation for a word or phrase in CMUdict (version 0.7b)

C M U Dictionary | Look It Up

☐ Show Lexical Stress

- C M U DICTIONARY
- S IY . EH M . Y UW . D IH K SH AH N EH R IY .


```
{
    "annotation": [
        {
            "path": "LibriSpeech/test-clean/61/70968/61-70968-0000.flac",
            "text": "sil HH IY1 B IH0 G AE1 N AH0 K AH0 N F Y UW1 Z D K AH0 M P L EY1 N T A
            "task": "phone_recognition"
        },
        {
            "path": "LibriSpeech/test-clean/61/70968/61-70968-0001.flac",
            "text": "sil G IH1 V N AA1 T S OW1 sil ER1 N IH0 S T AH0 M AY1 N D T AH0 DH IY1
            "task": "phone_recognition"
        },
        {
            "path": "LibriSpeech/test-clean/61/70968/61-70968-0002.flac",
            "text": "sil AH0 G OW1 L D AH0 N F AO1 R CH AH0 N AH0 N D AH0 HH AE1 P IY0 L AY
            "task": "phone_recognition"
        }
    ]
}
```

Overview – PR w/ non-strict post-processing

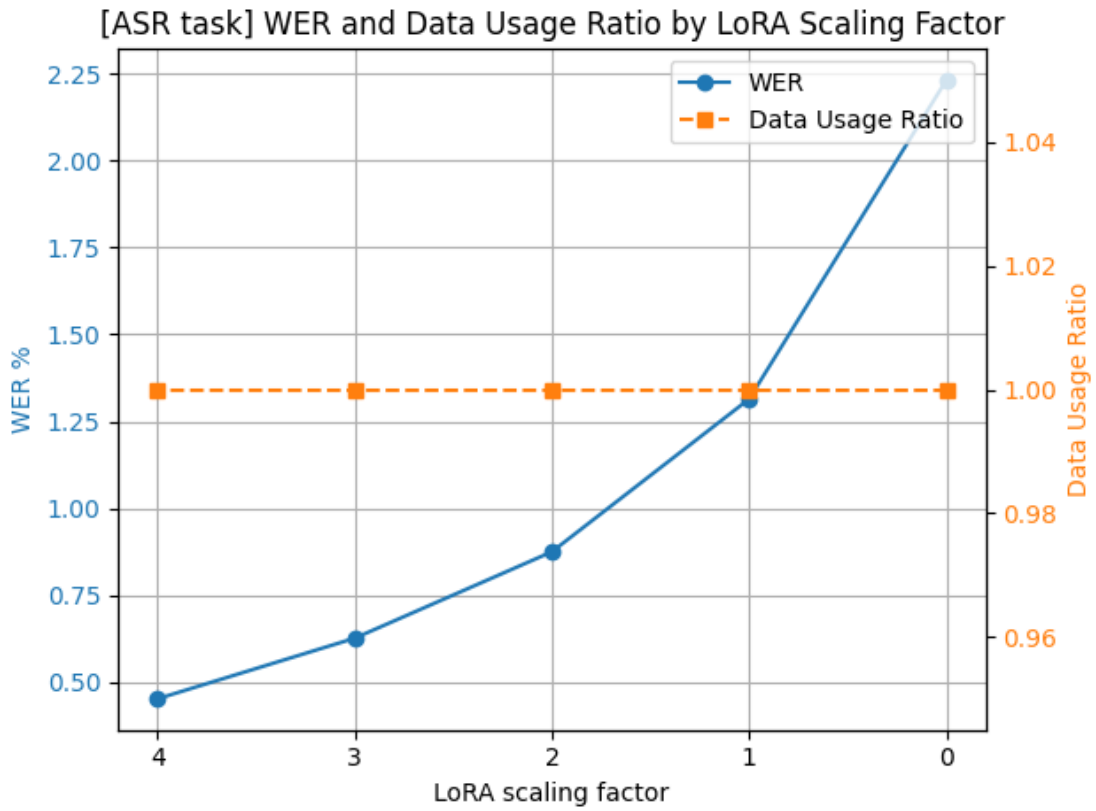
- LibriSpeech test-clean | 2,620
  - 0.04 (3.96%)

- note
  - WER

Method	PR↓
w/o Activation	4.2
w/ Activation	4.2
Reference Value	3.1

src: https://github.com/bytedance/SALMONN/issues/103

Figure3: Metrics by LoRA scaling factors | ASR w/o pp



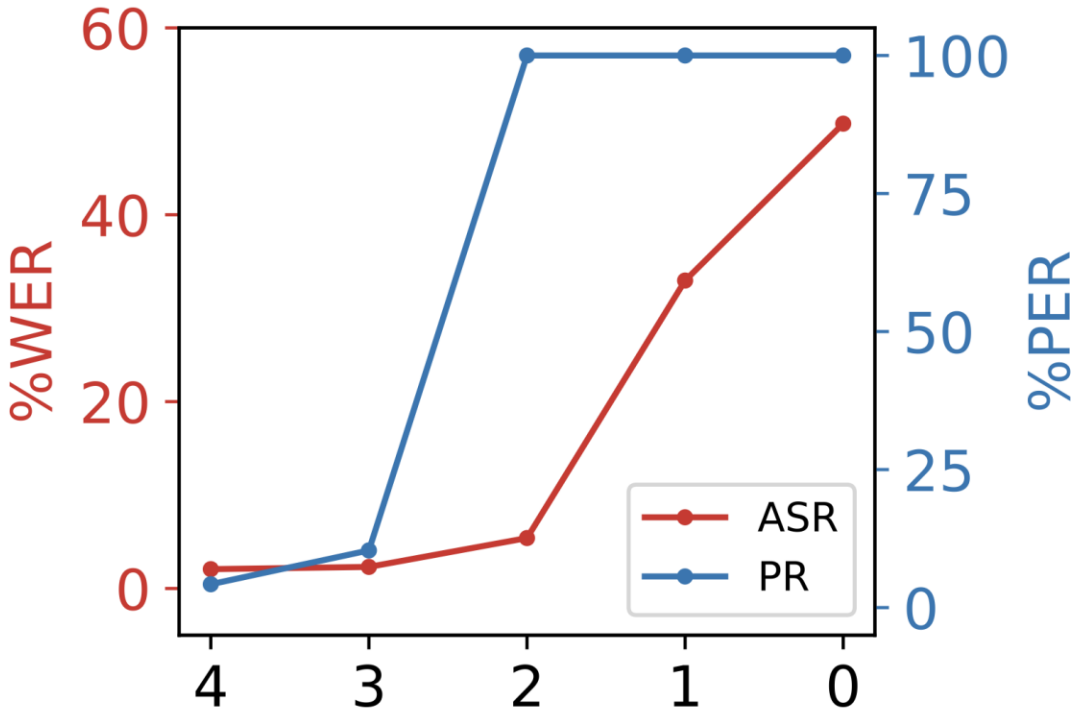
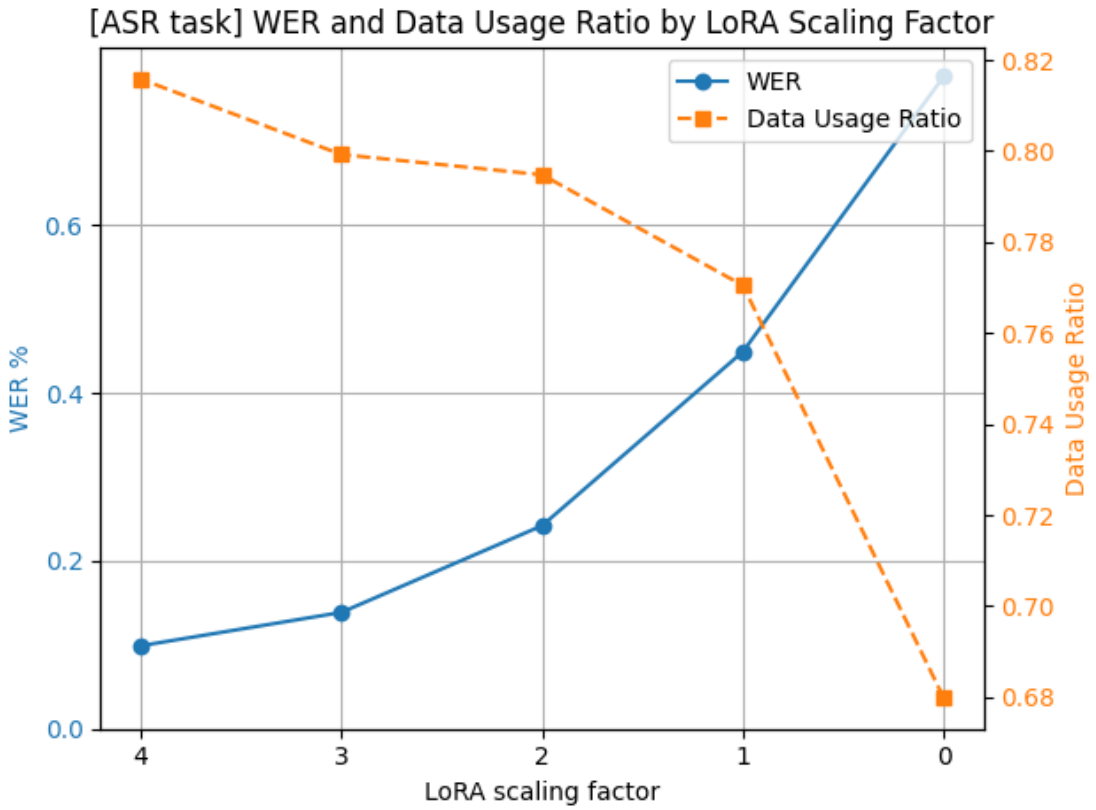


Figure3: Metrics by LoRA scaling factors | ASR w/ pp



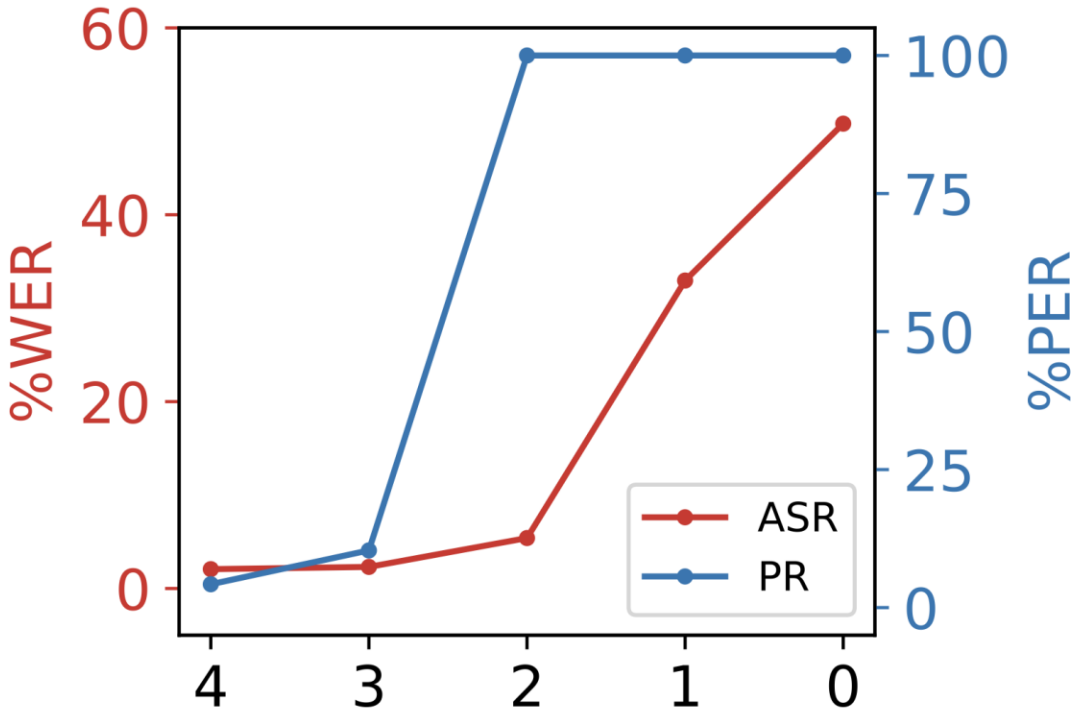
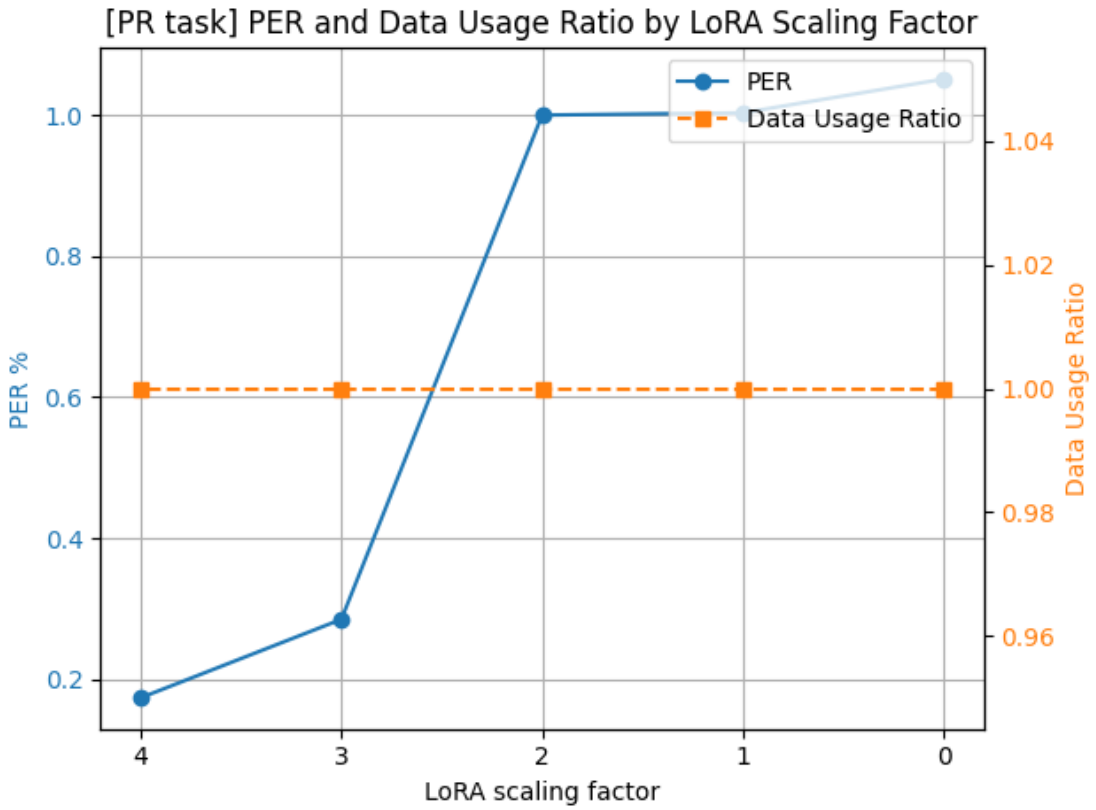


Figure3: Metrics by LoRA scaling factors | PR w/o pp



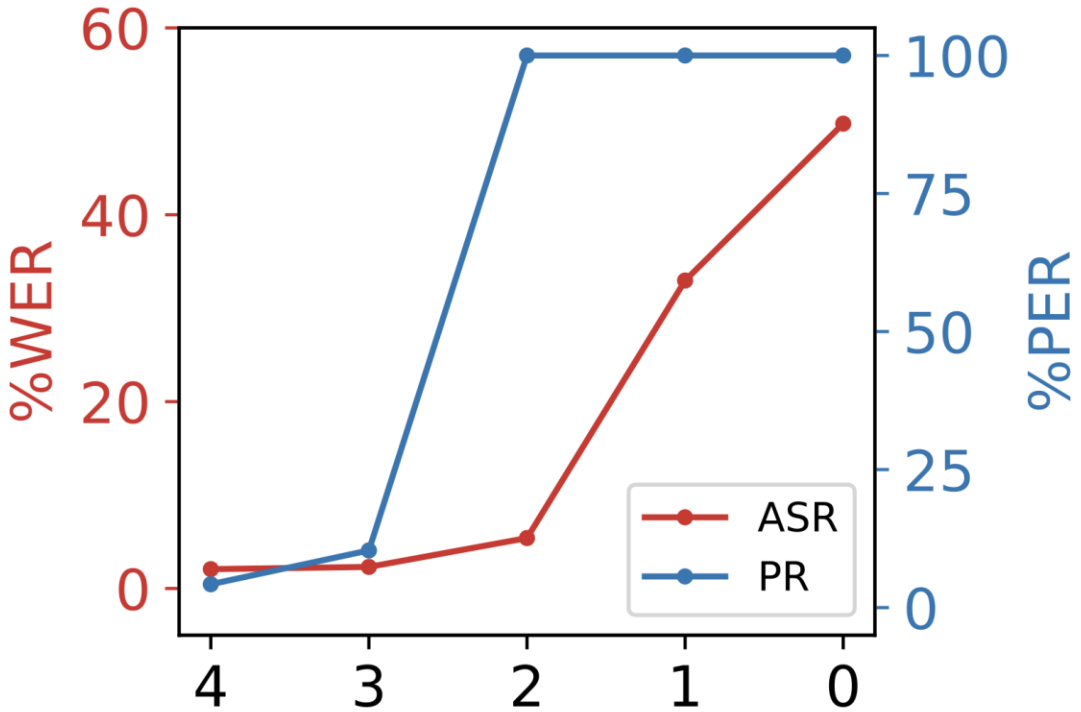
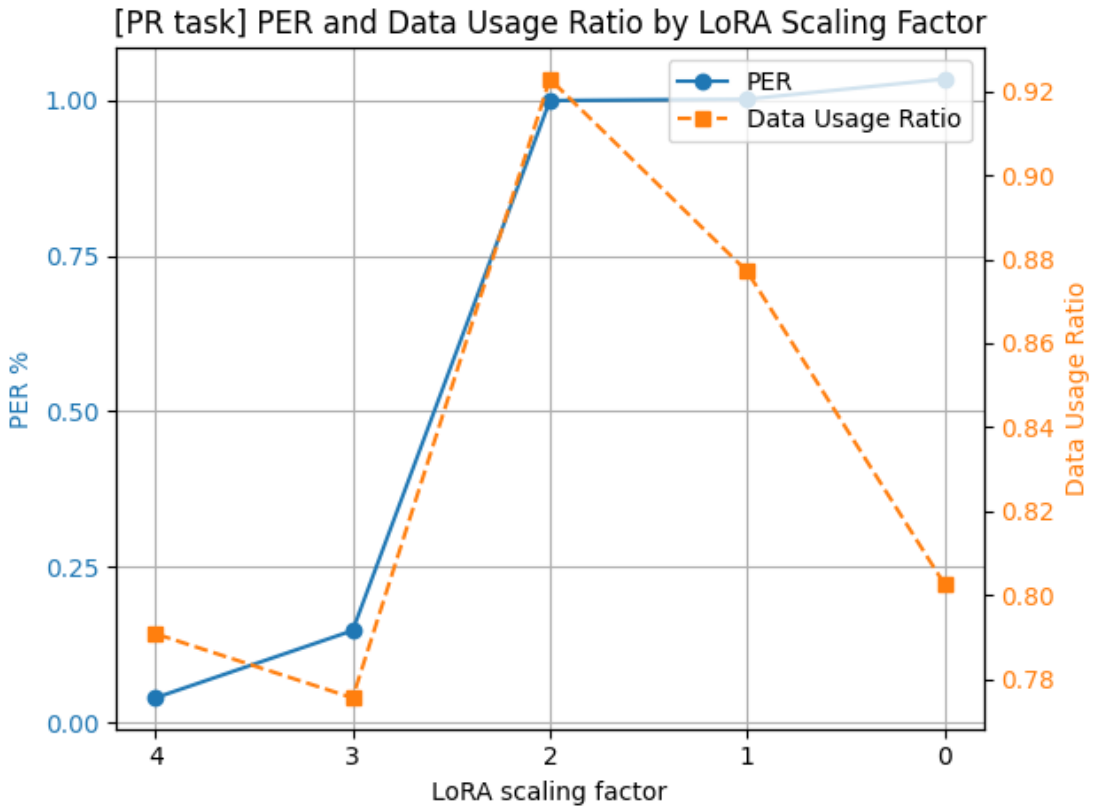
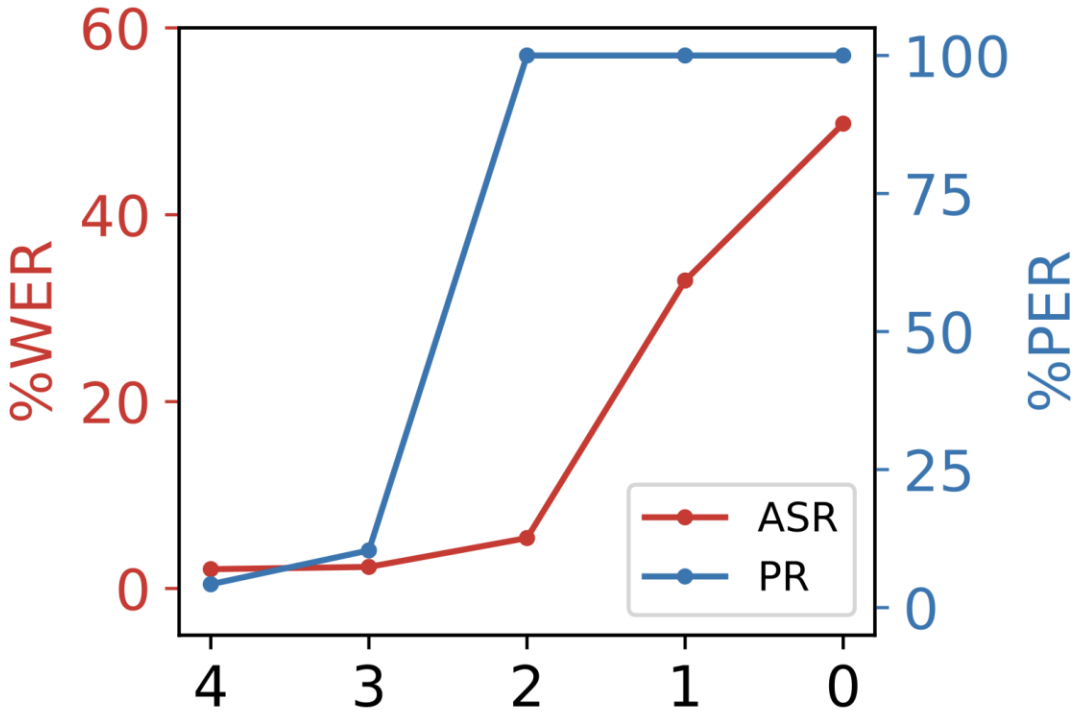


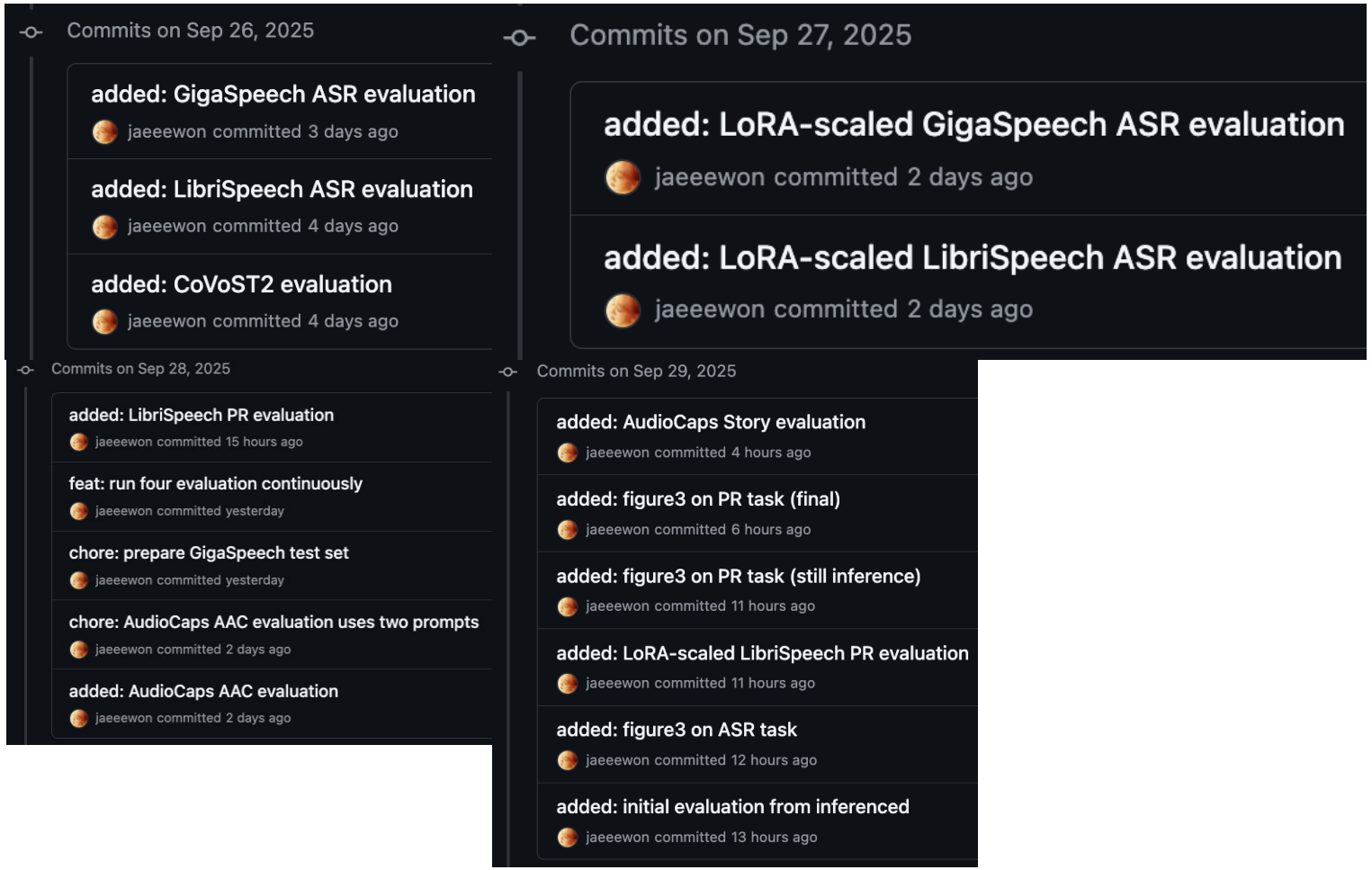
Figure3: Metrics by LoRA scaling factors | PR w/ pp





Conclusion

- postprocess is required to evaluate



DB0 | CoVoST2 – en2ja
DB1 | CoVoST2 – en2de
DB2 | LibriSpeech test-clean
DB3 | LibriSpeech test-other
DB4 | CoVoST2 – en2zh
DB5 | GigaSpeech test
DB6 | AudioCaps test

reproduce SALMONN

학생 원종찬, 최재원
Language & AI융합전공

2025. 09. 29. 3 PM.
한국외대 교수회관 401호

Post-Presentation-Summary

- Figure3: Metrics by LoRA scaling factors | ASR
  - they only used ‘LibriSpeech-ASR-test-clean’ for evaluation
  - postprocessing was not the key
    - https://github.com/jaeeewon/repr_salmonn/commit/722c7860ef0043649d45ec0b83bc543a780fc940