Self-Taught Recognizer: Toward Unsupervised Adaptation for Speech Foundation Models

Yuchen Hu*, Chen Chen*, Chao-Han Huck Yang, Chengwei Qin,
Pin-Yu Chen, Eng Siong Chng, Chao Zhang

NeurIPS 2024 (Poster)

Motivation

To deploy an ASR system in a practical scenario:

(NTU Canteen)

Collect

Supervised Domain Adaptation

Speech

Manual Labeling

+

Transcription

Neural Model

A very convenient approach is:

Unlabeled Speech

+

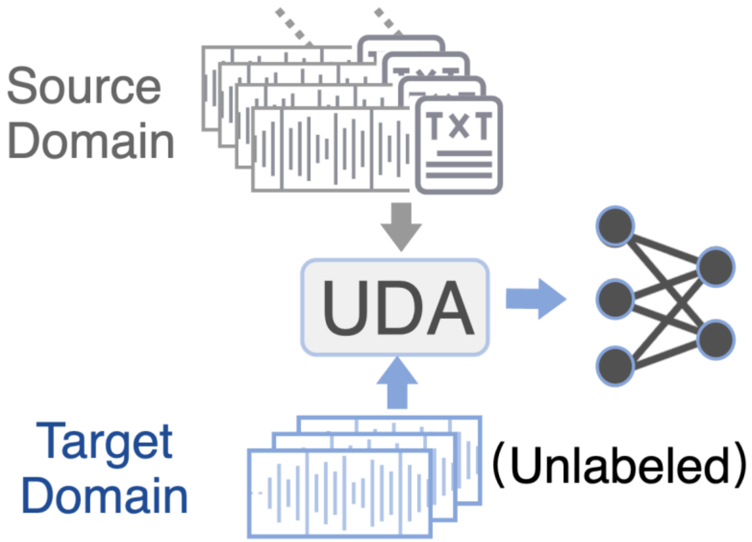
Whisper

Unsupervised Domain Adaptation

UDA

Motivation

UDA in ASR:

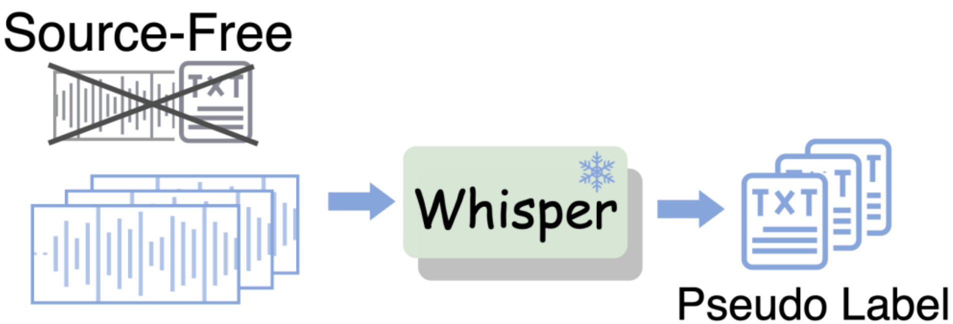


Human's UDA solution:

- "Unsupervised " is for adaptation process, but the learning schedule is semi-supervised.
- Considering the exhibited ability of large speech model:
  Can we skip the source-domain data for target domain adaptation? ➔ Source-free UDA

Method (Self-training/Semi-ASR)

1) Pseudo Labeling:

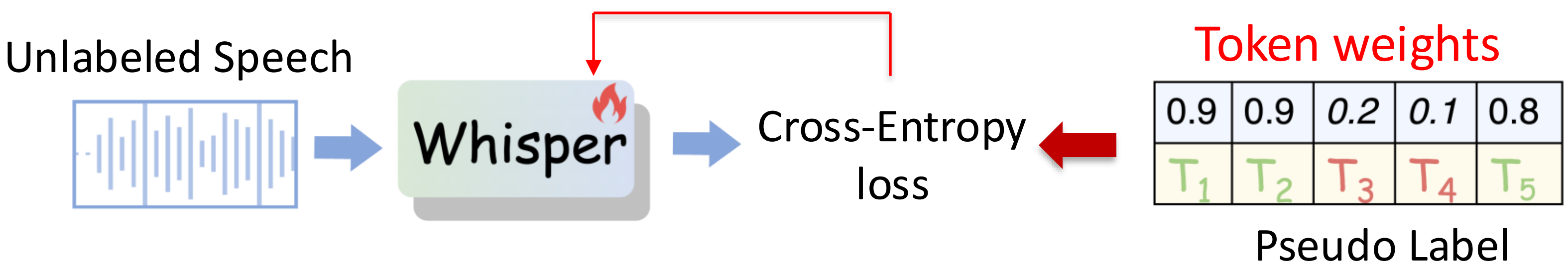


Keep this utterance or **discard**?

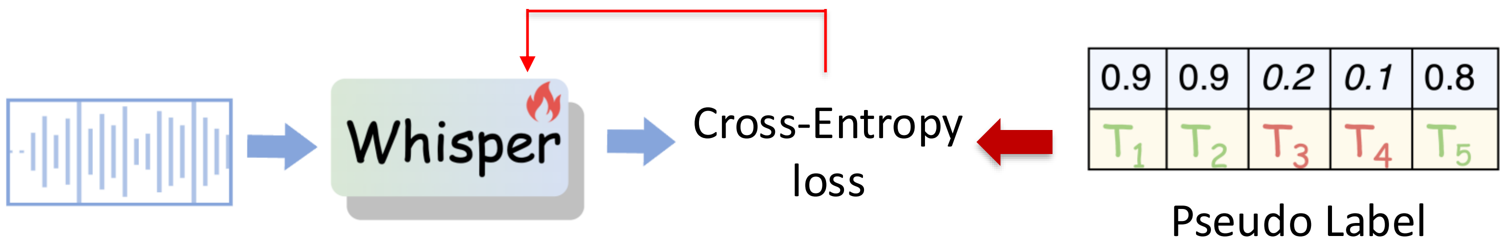
--> Monte Carlo sampling

2) Informed Finetuning:

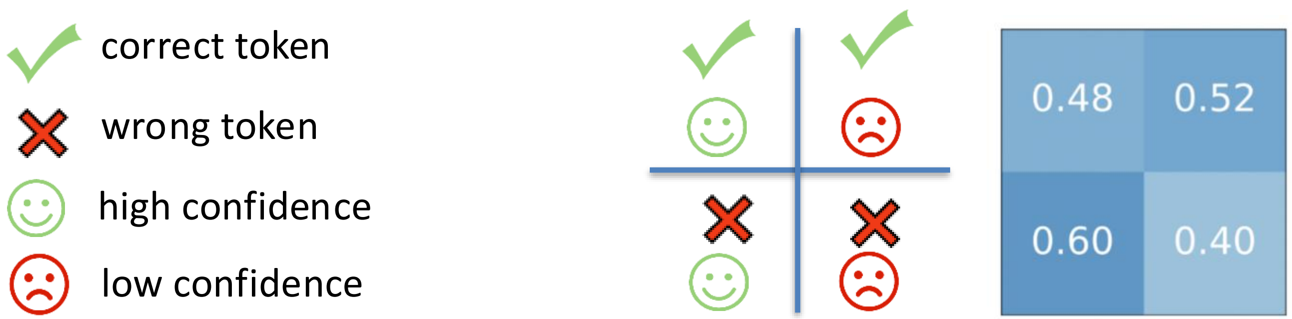
How to assign weight for **each token**?



Candidate 1: Confidence Score



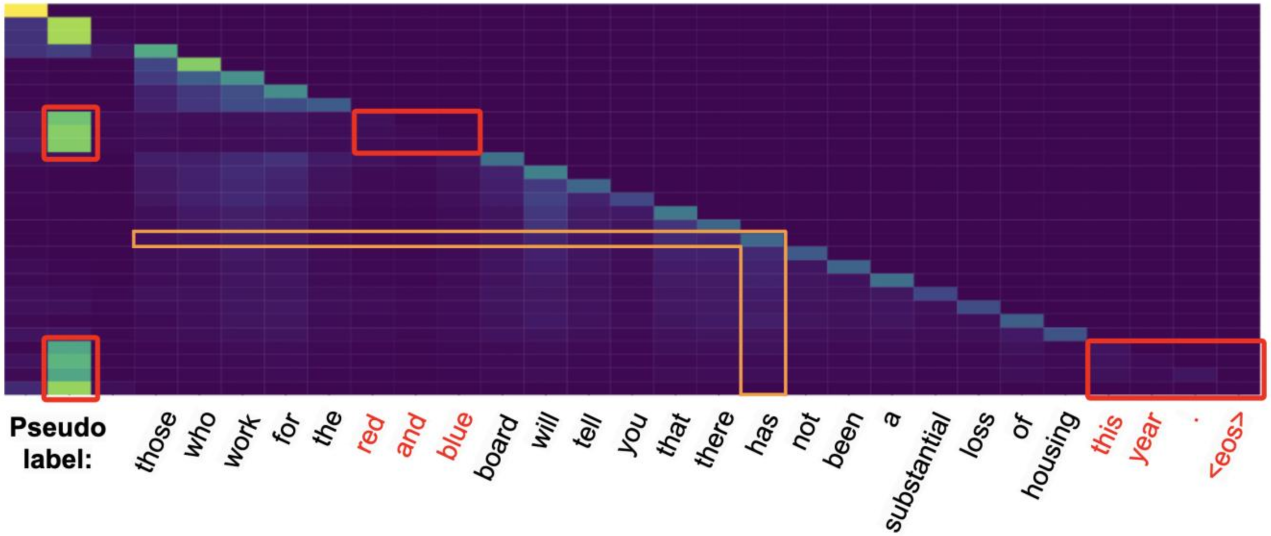
Experimental observation: decoding performance on CHiME-4 test-real



Confidence score is unreliable!

Candidate 2: Self-Attention Matrix

<|transcribe|>



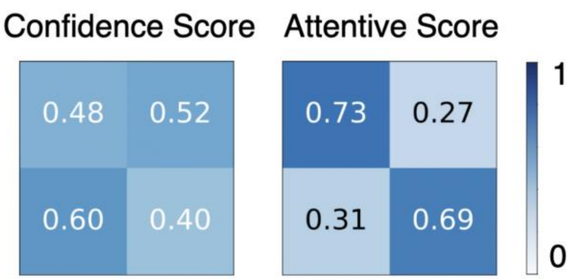
Attentive score:

$$\boxed{A_l = \sum_{j=4}^{l} W_{l,j} + \sum_{i=l+1}^{L} W_{i,l},}$$

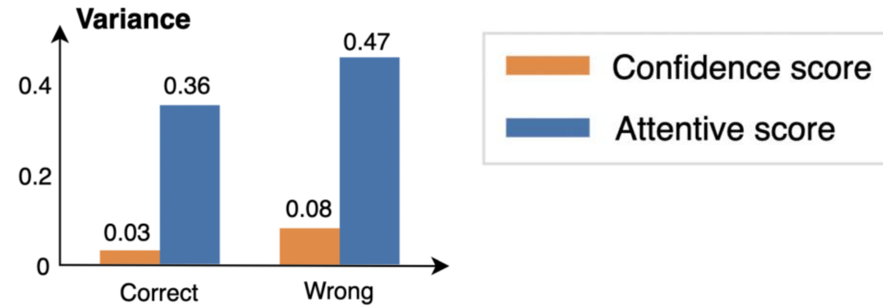
The importance of l-th token in whole utterance[8]

Correct/Wrong

Is A_l more ***reliable*** *than C_l ?*



Is A_l stable for guide finetuning?



Conclusion: attentive score is **more reliable** but **less stable** than confidence score.

STAR: Integrate *A* and *C* for each token

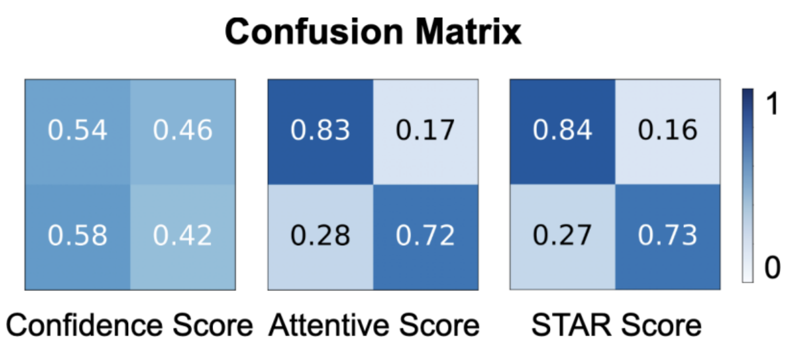
Criteria: - If *A*-*C* conflict, then follow *A:*

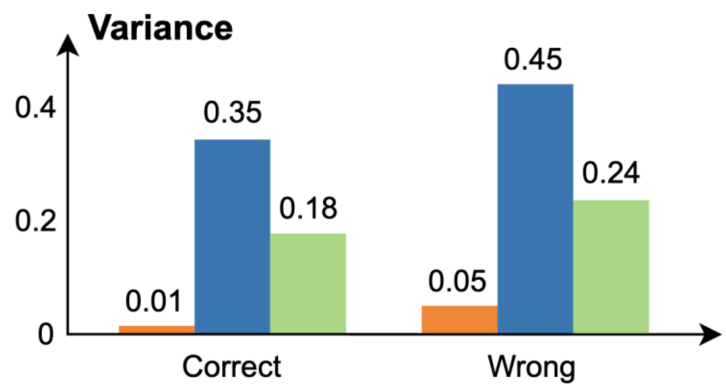
$$\mathcal{S}_l^{\text{conf}} = [\sigma(\mathcal{A}_l^2/\mathcal{C}_l - \lambda) + \sigma(\mathcal{C}_l^2/\mathcal{A}_l - \lambda)] * \mathcal{A}_l$$

- If *A*-*C* consistent, then calibrate *A* using *C:*

$$\mathcal{S}_l^{\text{cons}} = [\sigma(\lambda - \mathcal{A}_l^2/\mathcal{C}_l) * \sigma(\lambda - \mathcal{C}_l^2/\mathcal{A}_l)] * \mathcal{A}_l * e^{(\mathcal{C}_l - \mathcal{A}_l)/\tau}.$$

Quick validation:





Effectiveness on Various Domains

STAR = Self-TAught Recognizer

Testing Scenario		Whisper (frozen)	Whisper (self-train.)	$\text{UTT}_{\text{filter}}$	$\text{TOK}_{\text{reweight}}$ $\mathcal{C}_l$	$\mathcal{A}_l$	**STAR (ours)**	Whisper (real label)
			Background Noise					
CHiME-4	*test-real*	6.8	6.9	6.4	6.5	6.2	**6.0** $_{-11.8\%}$	5.2
	test-simu	9.9	10.1	9.7	9.8	9.5	**9.4** $_{-5.1\%}$	8.7
	dev-real	4.6	4.5	4.3	4.3	4.1	**3.9** $_{-15.2\%}$	3.2
	dev-simu	7.0	7.0	6.6	6.7	6.6	**6.4** $_{-8.6\%}$	5.9
LS-FreeSound	*babble*	40.2	37.6	35.0	33.5	31.3	**30.2** $_{-24.9\%}$	27.2
	airport	15.6	15.5	15.2	15.3	15.0	**14.8** $_{-5.1\%}$	14.5
	car	2.9	3.0	2.8	2.8	2.6	**2.5** $_{-13.8\%}$	2.4
RATS	*radio*	46.9	47.2	46.0	45.5	44.9	**44.6** $_{-4.9\%}$	38.6
			Speaker Accents					
CommonVoice	*African*	6.0	5.8	5.5	5.4	5.0	**4.8** $_{-20.0\%}$	4.6
	Australian	5.8	5.7	5.6	5.5	5.2	**5.1** $_{-12.1\%}$	4.3
	Indian	6.6	6.5	6.3	6.4	6.1	**6.0** $_{-9.1\%}$	5.7
	Singaporean	6.5	6.2	5.8	5.8	5.4	**5.1** $_{-21.5\%}$	4.9
			Specific Scenarios					
TED-LIUM 3	*TED talks*	5.2	4.9	4.7	4.8	4.3	**4.1** $_{-21.2\%}$	3.6
SwitchBoard	*telephone*	20.8	20.5	19.8	19.3	18.6	**18.1** $_{-13.0\%}$	15.3
LRS2	*BBC talks*	8.5	8.3	7.6	7.9	7.4	**7.0** $_{-17.6\%}$	5.6
ATIS	*airline info.*	3.6	3.5	3.3	3.3	3.2	**2.9** $_{-19.4\%}$	2.0
CORAAL	*interview*	21.5	21.3	20.8	20.7	20.4	**20.1** $_{-6.5\%}$	17.9

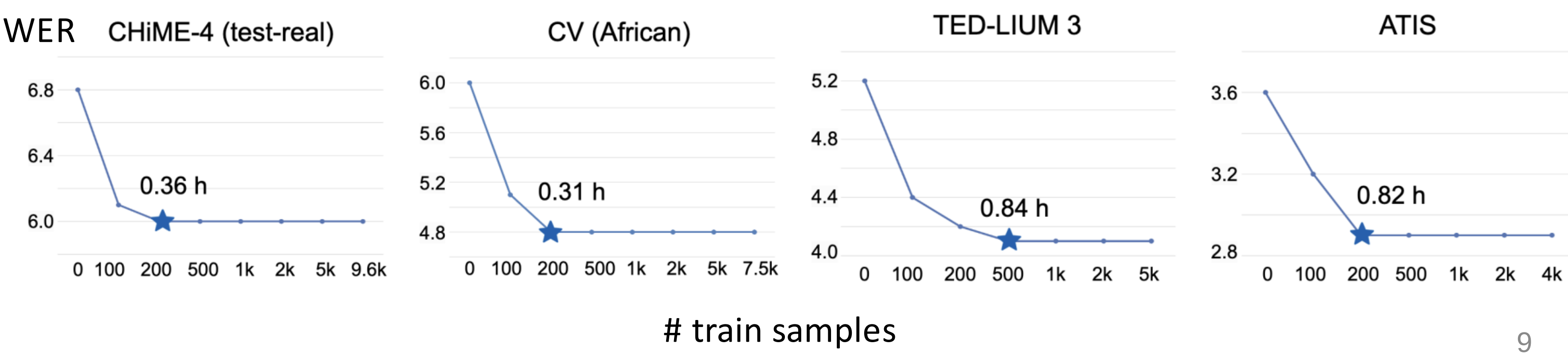
Analysis

STAR can avoid forgetting:

Model	LS-FreeSound			RATS	CommonVoice				TED-3	SWBD	ATIS
	babble	*airport*	*car*		*af*	*au*	*in*	*sg*			
Frozen	40.2	**15.6**	2.9	46.9	**6.0**	**5.8**	**6.6**	6.5	5.2	**13.3**	3.6
Self-train.	38.2	16.6	2.9	47.3	6.4	5.9	6.7	6.3	5.3	13.7	3.4
STAR	**33.3**	15.7	**2.8**	**46.1**	6.1	**5.8**	6.7	**5.6**	**5.0**	13.5	**2.9**

Train on CHiME-4 and test on OOD

STAR enjoys high data efficiency:



Generalization

- Other models

Model	Baseline	Self-train.	STAR	Real
Whisper-V3-1.5B	6.8	6.9	$6.0_{-11.8\%}$	5.2
Whisper-Med-0.8B	8.9	8.8	$8.0_{-10.1\%}$	7.1
OWSM-V3.1-1.0B	8.4	8.1	$7.5_{-10.7\%}$	6.5
Canary-1.0B	8.2	8.0	$7.2_{-12.2\%}$	6.4
Parakeet-TDT-1.1B	8.0	7.8	$7.0_{-12.5\%}$	6.2

- Other task (Speech Translation on FLURS)

$X \rightarrow$ En	Baseline	Self-train.	STAR	Real
Ar	21.9	22.1	$23.3_{+1.4}$	24.5
De	33.7	34.0	$35.9_{+2.2}$	36.5
Es	23.9	24.1	$24.8_{+0.9}$	26.4
Fa	16.6	16.3	$17.6_{+1.0}$	19.0
Hi	22.4	22.5	$23.4_{+1.0}$	24.4
Zh	16.3	16.3	$17.1_{+0.8}$	17.9

Ablation Study

- Different whisper sizes

Model Size	# Param.	Baseline	STAR	Real
large-v3	1,550 M	6.8	$6.0_{-11.8\%}$	5.2
large-v2		7.7	$6.9_{-10.4\%}$	6.0
large		7.5	$7.0_{-6.7\%}$	6.8
medium.en	769 M	8.9	$8.0_{-10.1\%}$	7.1
small.en	244 M	12.7	$10.6_{-16.5\%}$	9.0
base.en	74 M	32.4	$17.7_{-45.4\%}$	16.1

- Different training methods

Approach	# Param.*	Baseline	STAR	Real
Regular Finetuning				
Full	1550 M	6.8	$6.0_{-11.8\%}$	5.2
Enc-only	635 M		$6.3_{-7.4\%}$	5.0
Dec-only	907 M		$6.1_{-10.3\%}$	4.4
Parameter-Efficient Finetuning				
LoRA	16 M	6.8	$6.0_{-11.8\%}$	5.1
Reprogram.	0.4 M		$6.7_{-1.5\%}$	6.7

Iterative Finetuning

Model	Test set	# Iterations 0	1	2	3	4	5	Real label
large-v3	*test-real*	6.8	6.0	5.9	5.7	5.7	5.7	5.2
medium.en	*test-real*	8.9	8.0	7.9	7.9	7.8	7.8	7.1
small.en	*test-real*	12.7	10.6	10.3	10.3	10.3	10.3	9.0
base.en	*test-real*	34.4	17.7	17.2	17.2	17.0	17.0	16.1
large-v3	*test-simu*	9.9	9.4	9.3	9.0	8.9	8.9	8.7
large-v3	*dev-real*	4.6	3.9	3.9	3.8	3.8	3.8	3.2
large-v3	*dev-simu*	7.0	6.4	6.4	6.4	6.3	6.3	5.9
large-v3	*af*	6.0	4.8	4.8	4.7	4.7	4.7	4.6
large-v3	*au*	5.8	5.1	5.0	4.6	4.5	4.5	4.3
large-v3	*in*	6.6	6.0	5.8	5.8	5.8	5.8	5.7
large-v3	*sg*	6.5	5.1	5.1	5.1	5.1	5.1	4.9

- Iterative post-training can further improve results
- Little further improvements after 3 iterations

Conclusion & Discussion

Easy-to-use:

- A pretrained Model + 1-hour ***unlabeled*** speech
- **13.5%** relative WER reduction across **14** target domains (noise, accent, etc.)

Generalization:

- Other models: SeamlessM4T, OWSM, Canary
- Other task: Speech Translation

Anti-forgetting:

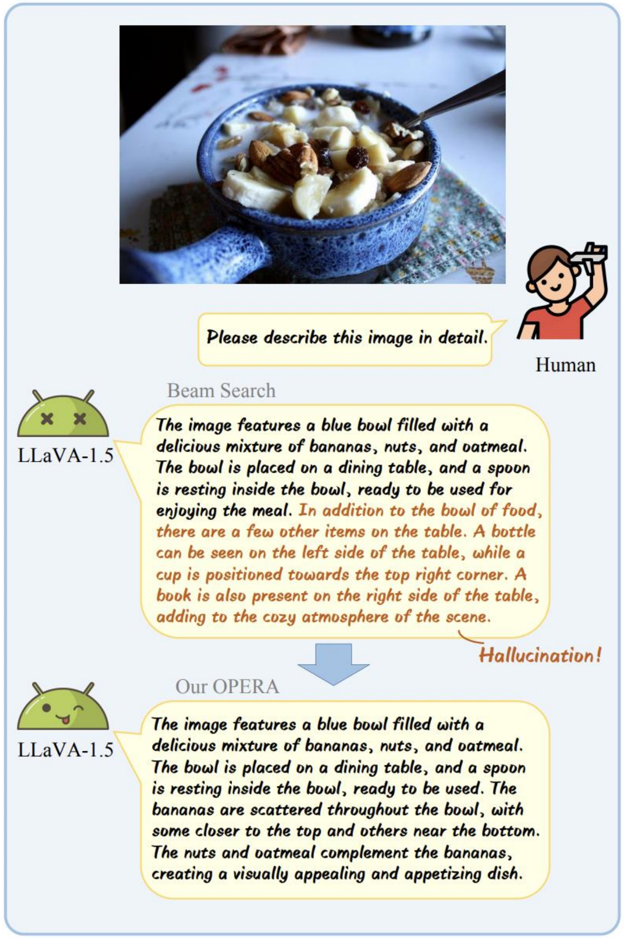
- Avoid common catastrophic forgetting in domain adaptation

Discussion

- Large models' attention matrix can present their uncertainty
- Self-improvement is possible in large speech foundation Model

Thank you! & QA

Appendix: LLM Hallucination



- **Non-Hallucinations:** describes the food (e.g., bananas, nuts, oatmeal) inside the bowel
- **Hallucinations:** imagines the items on the table that is outside the image

***NOTE:** Hallucinations starts with "In addition to …"*

Appendix: LLM Hallucination

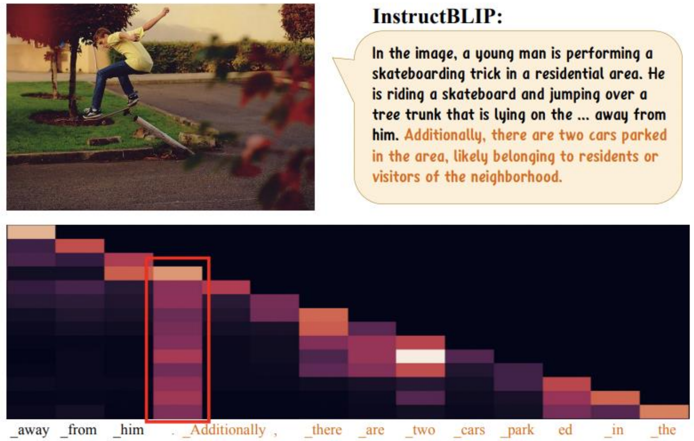


Figure 2. A case of relationship between hallucinations and knowledge aggregation patterns. Hallucinations are highlighted.

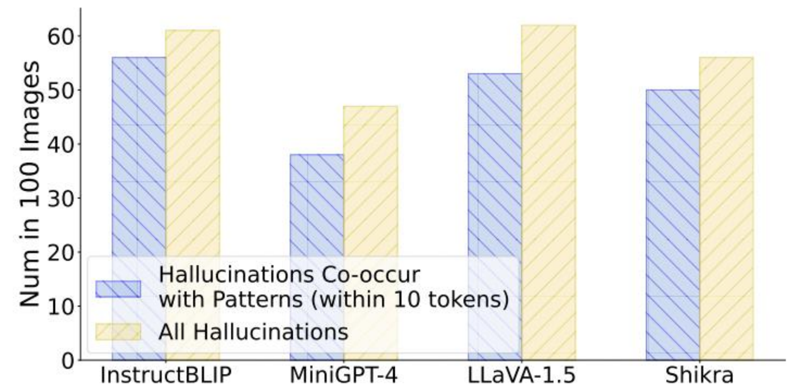


Figure 3. Hallucinations often start within the first 10 tokens after knowledge aggregation patterns.

- Hallucinations are usually triggered by specific tokens (e.g., “*additionally*”);
- We can observe a “knowledge aggregation pattern” in self-attention map along with the beginning of hallucinations → *An insightful finding!*

Appendix: LLM Hallucination

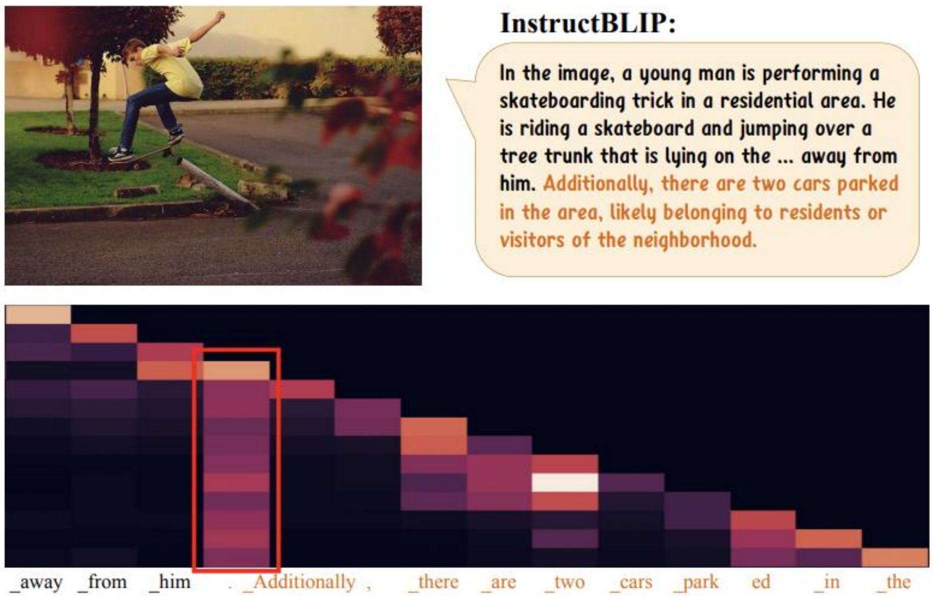


Figure 2. A case of relationship between hallucinations and knowledge aggregation patterns. Hallucinations are highlighted.

All hallucinations are highly related to the starting token "*additionally*" but unrelated to previous normal tokens!