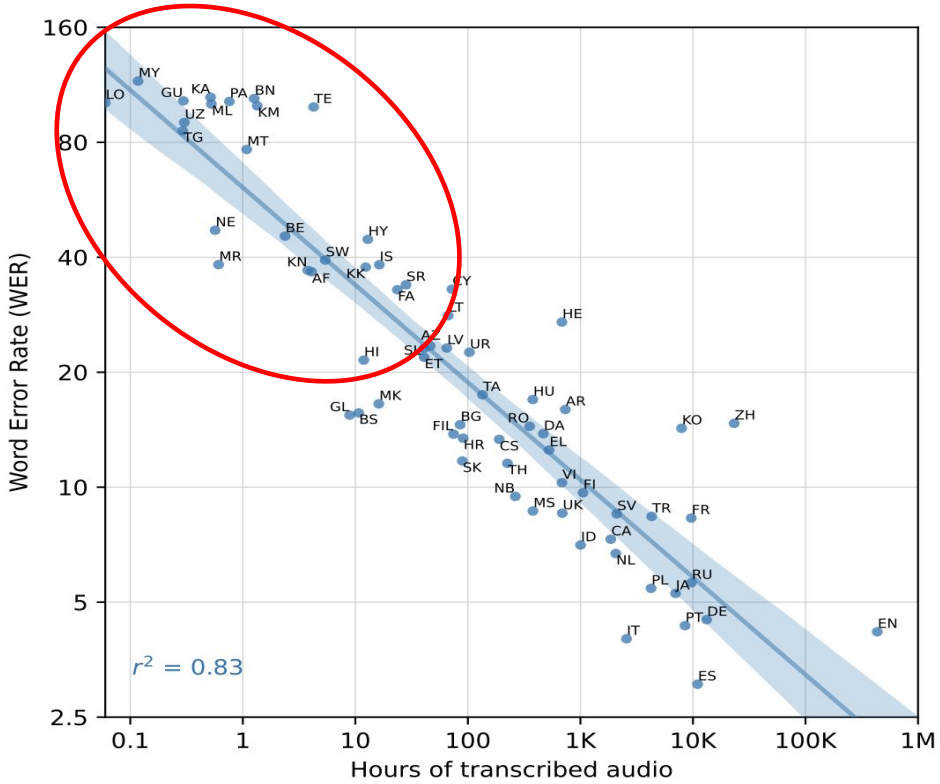
SALSA: Speedy ASR-LLM Synchronous Aggregation

Ashish Mittal[1,2], Darshan Prabhu[2],
Sunita Sarawagi[2], Preethi Jyothi[2]

[1]IBM Research, [2]IIT Bombay

Accepted at Interspeech 2024

ASR performance



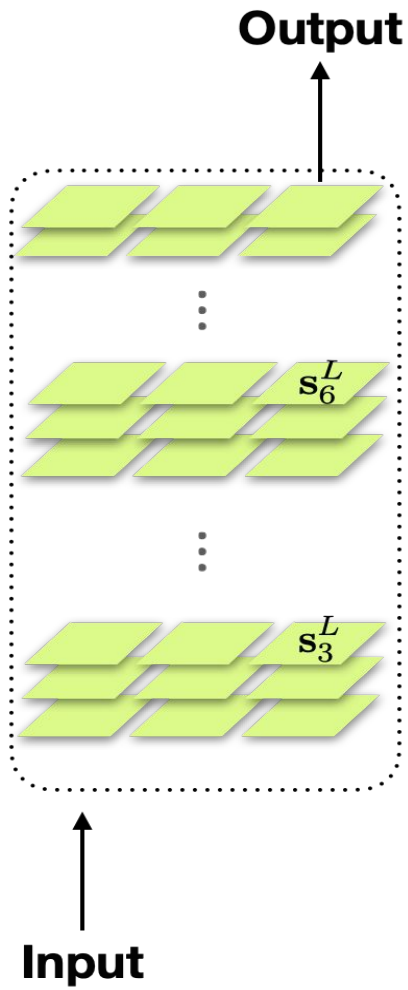
High-Resource Languages: Decent Performance

Low-Resource Languages:

- Performance remains suboptimal
- Acquiring transcribed is challenging.

Reference: Radford, Alec, et al. "Robust speech recognition via large-scale weak supervision.

Large Language Models



Large Language Model (LLM)

- Massive scale
- Impressive language generation
- Multilingual generation

ASR-LLM Integration

- Beneficial for low-resource languages for which LLM has reasonable base capabilities.

ASR-LLM Integration

- Beneficial for low-resource languages for which LLM has reasonable base capabilities.

- Prior LLM-ASR integration techniques are expensive
  - High training overhead (e.g., speech in-context learning)
  - High decoding latency (e.g., second pass rescoring in ASR error correction).

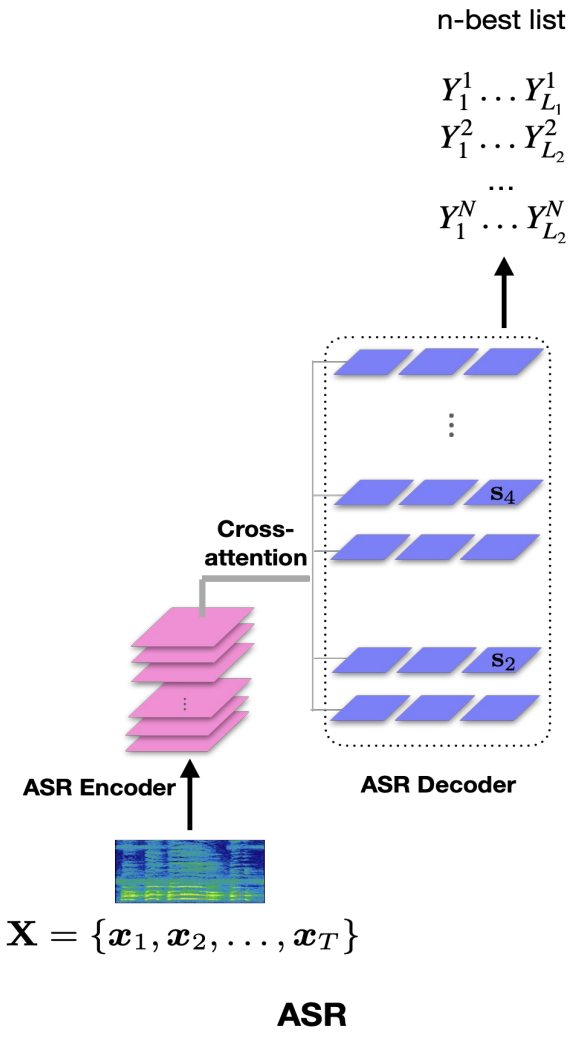
ASR-LLM Integration

- Beneficial for low-resource languages for which LLM has reasonable base capabilities.

- Prior LLM-ASR integration techniques are expensive
  - High training overhead (e.g., speech in-context learning)
  - High decoding latency (e.g., second pass rescoring in ASR error correction).

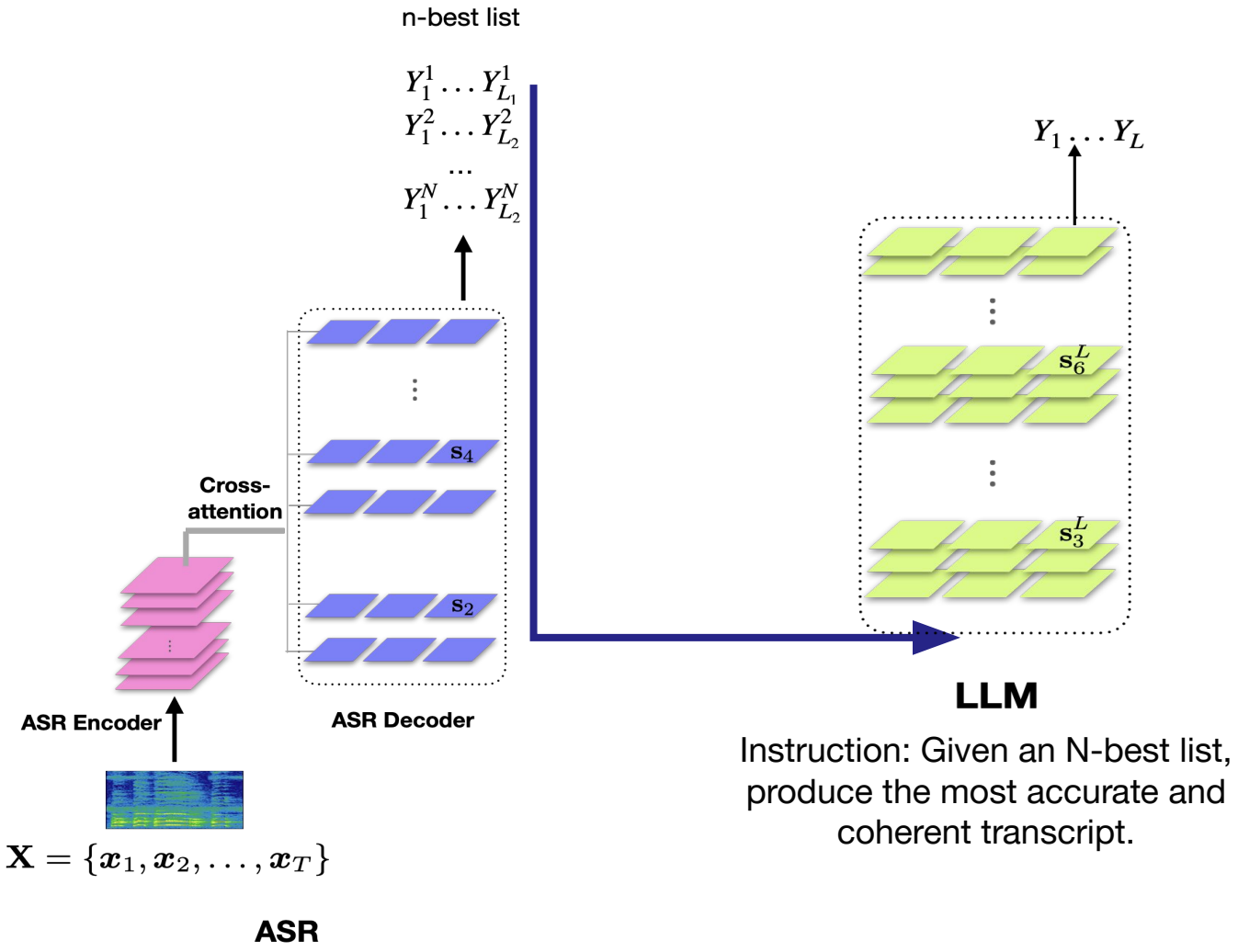
- We propose -- a lightweight approach SALSA that couples the decoder layers of pretrained ASR and LLM models.

N-best List



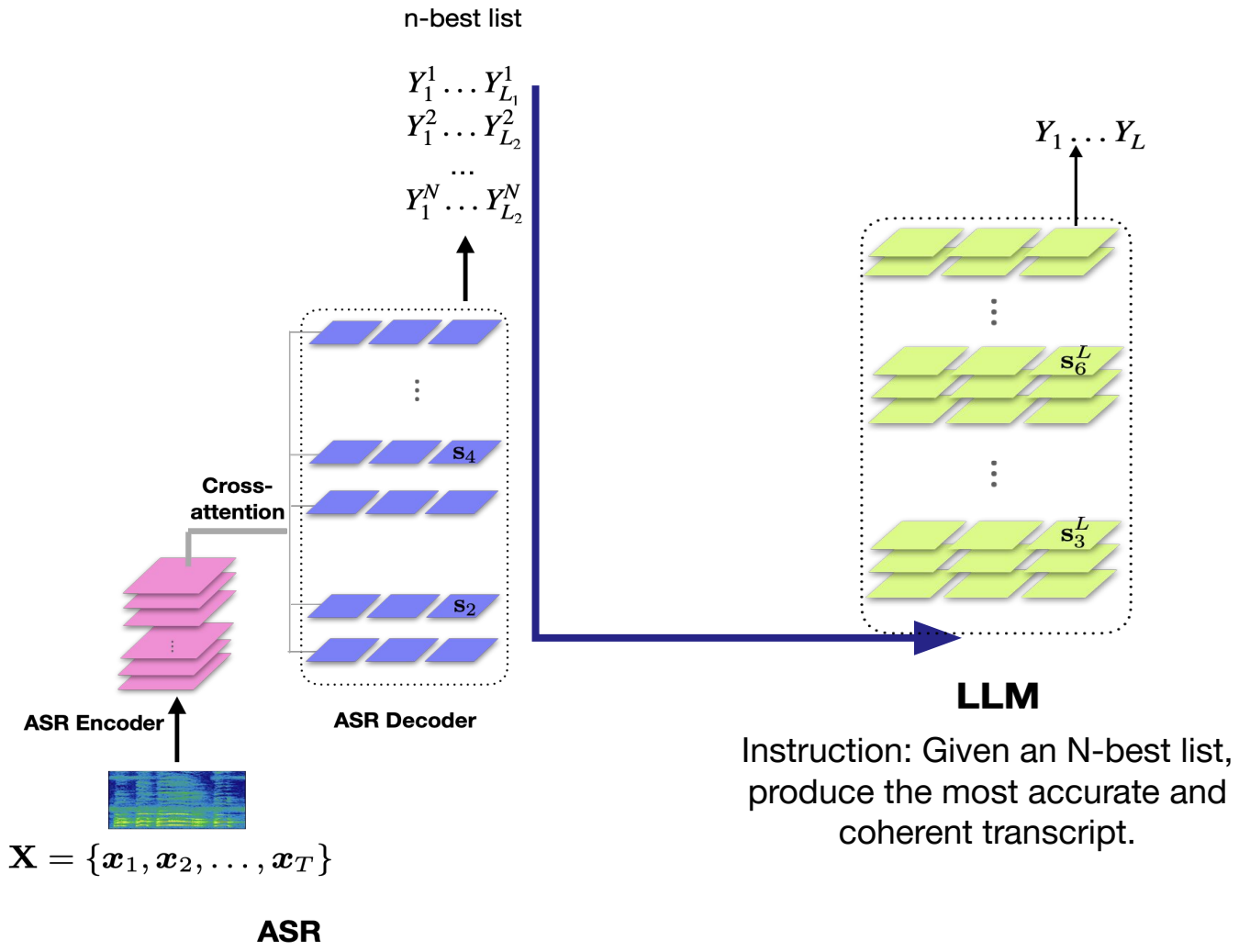
References:
[1] C.-H. H. Yang, Y. Gu, Y.-C. Liu, S. Ghosh, I. Bulyko, and A. Stolcke, "" Generative speech recognition error correction with large language models and task-activating prompting
[2] P. Dighe, Y. Su, S. Zheng, Y. Liu, V. Garg, X. Niu, and A. Tewfik, "Leveraging large language models for exploiting asr uncertainty"
[3] R. Ma, M. Qian, P. Manakul, M. Gales, and K. Knill, "Can generative large language models perform asr error correction?"

N-best List



References:
[1] C.-H. H. Yang, Y. Gu, Y.-C. Liu, S. Ghosh, I. Bulyko, and A. Stolcke, "" Generative speech recognition error correction with large language models and task-activating prompting
[2] P. Dighe, Y. Su, S. Zheng, Y. Liu, V. Garg, X. Niu, and A. Tewfik, "Leveraging large language models for exploiting asr uncertainty"
[3] R. Ma, M. Qian, P. Manakul, M. Gales, and K. Knill, "Can generative large language models perform asr error correction?"

N-best List

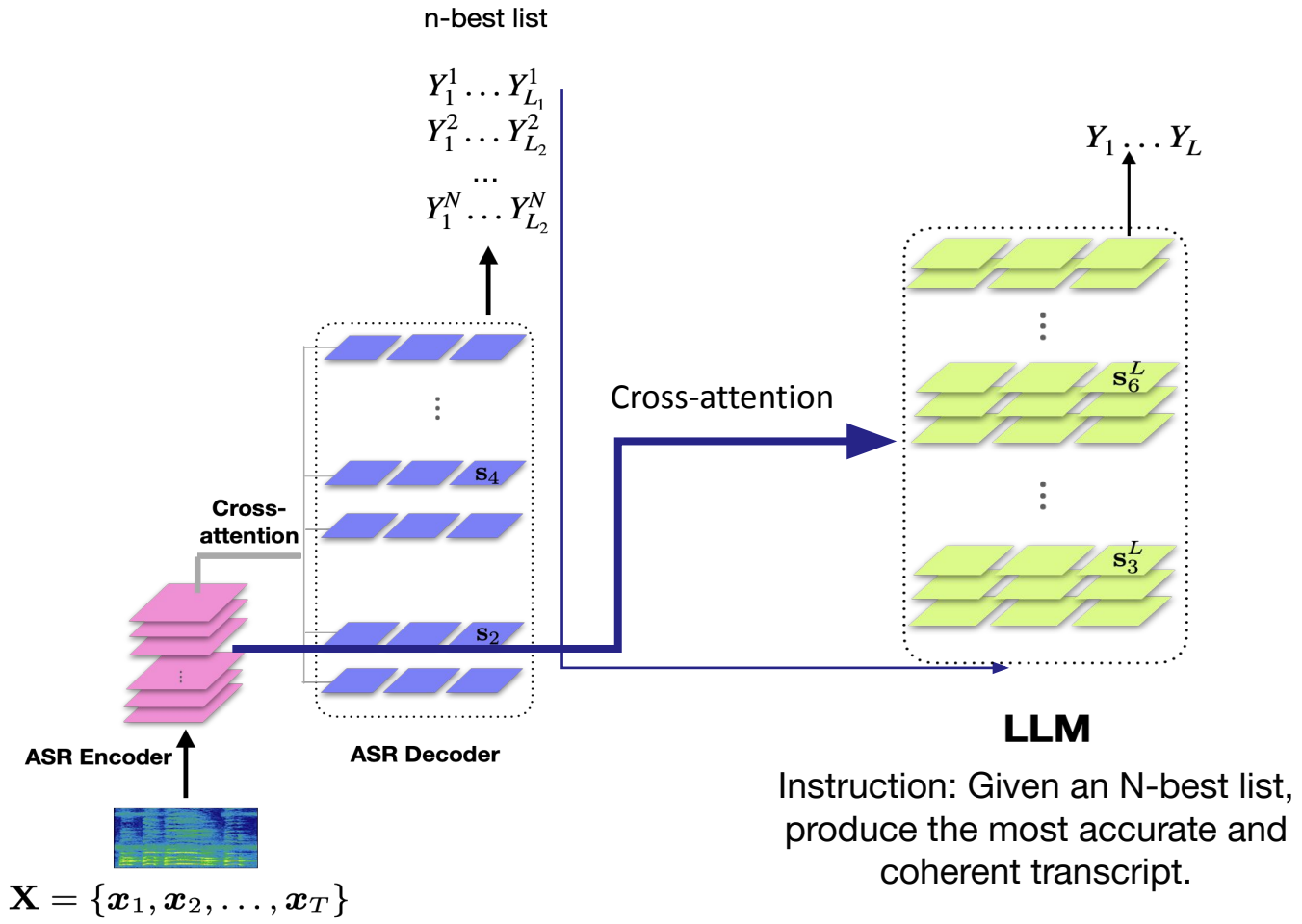


Limitations:

- N-best lists unreliable for low resource languages
- High Latency
- LLM agnostic to audio

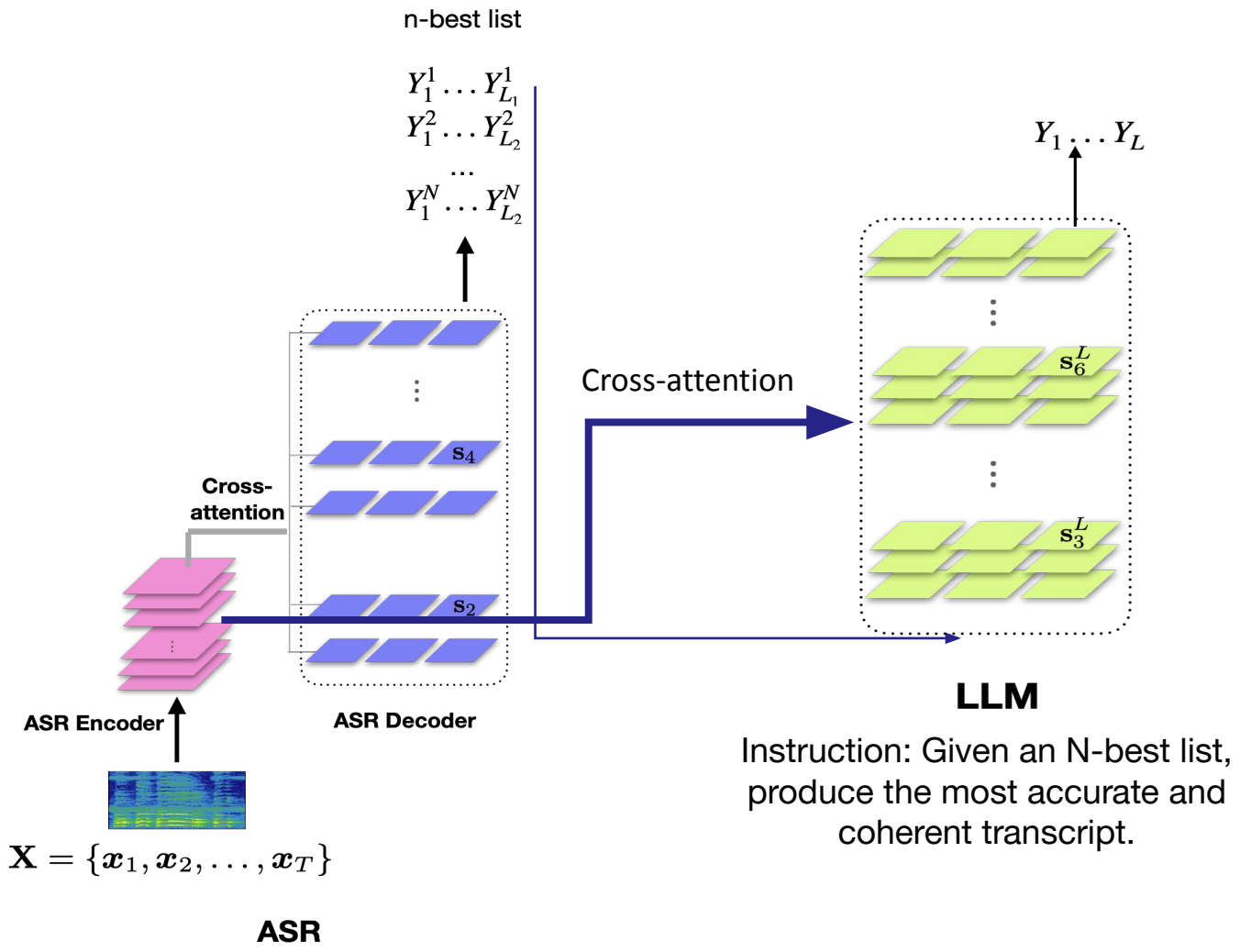
References:
[1] C.-H. H. Yang, Y. Gu, Y.-C. Liu, S. Ghosh, I. Bulyko, and A. Stolcke, "" Generative speech recognition error correction with large language models and task-activating prompting
[2] P. Dighe, Y. Su, S. Zheng, Y. Liu, V. Garg, X. Niu, and A. Tewfik, "Leveraging large language models for exploiting asr uncertainty"
[3] R. Ma, M. Qian, P. Manakul, M. Gales, and K. Knill, "Can generative large language models perform asr error correction?"

Whispering-Llama (Radhakrishnan et al.)



References:
[1] C.-H. H. Yang, Y. Gu, Y.-C. Liu, S. Ghosh, I. Bulyko, and A. Stolcke, "Generative speech recognition error correction with large language models and task-activating prompting"
[2] P. Dighe, Y. Su, S. Zheng, Y. Liu, V. Garg, X. Niu, and A. Tewfik, "Leveraging large language models for exploiting asr uncertainty"
[3] R. Ma, M. Qian, P. Manakul, M. Gales, and K. Knill, "Can generative large language models perform asr error correction?"
[4] S. Radhakrishnan et al., "Whispering llama: A cross-modal generative error correction framework for speech recognition"

Whispering-Llama (Radhakrishnan et al.)

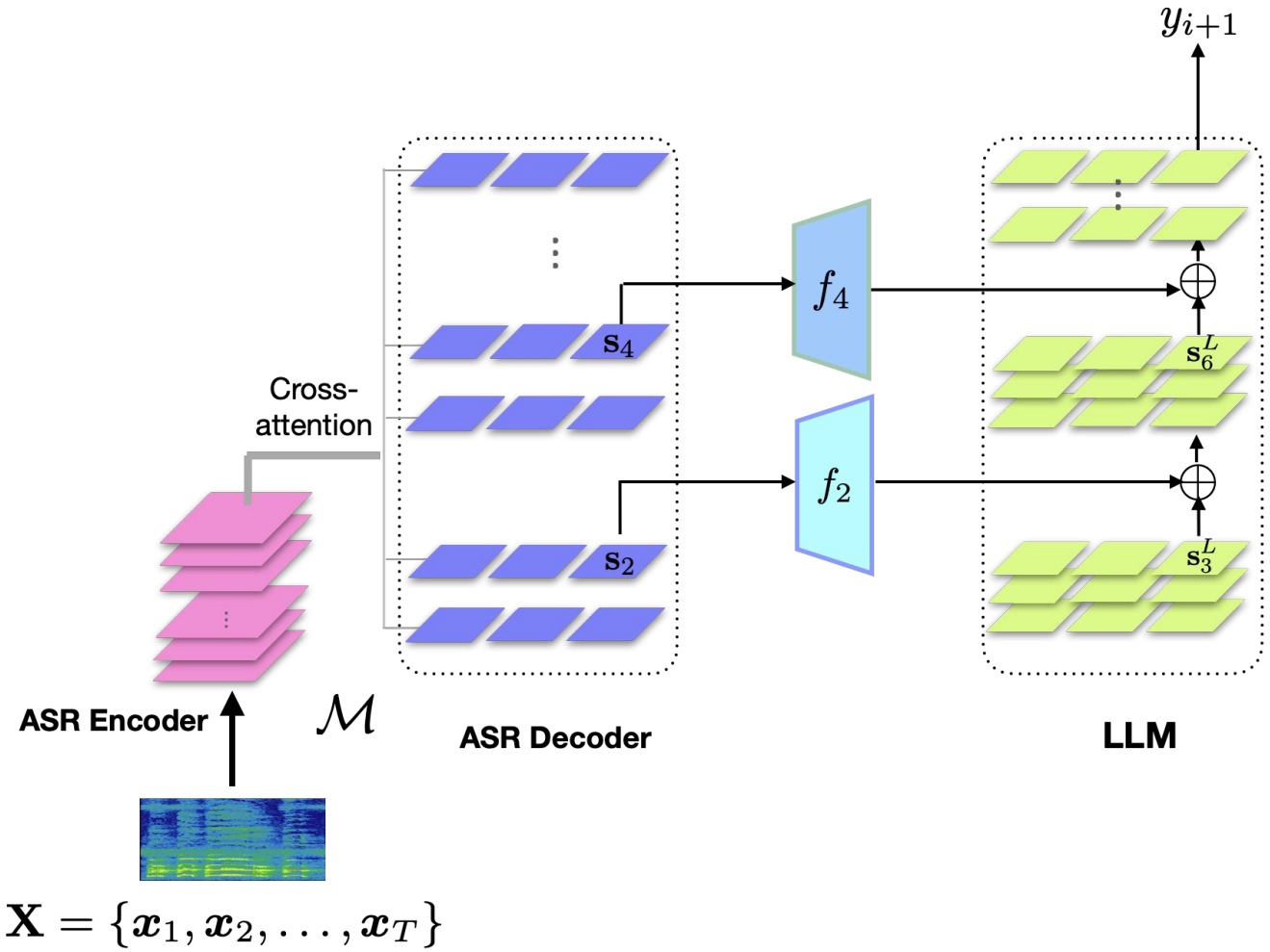


Limitations:

- N-best lists unreliable for low resource languages
- High Latency
- ~~LLM agnostic to audio~~
- Large training data to learn cross-attention

References:
[1] C.-H. H. Yang, Y. Gu, Y.-C. Liu, S. Ghosh, I. Bulyko, and A. Stolcke, "Generative speech recognition error correction with large language models and task-activating prompting"
[2] P. Dighe, Y. Su, S. Zheng, Y. Liu, V. Garg, X. Niu, and A. Tewfik, "Leveraging large language models for exploiting asr uncertainty"
[3] R. Ma, M. Qian, P. Manakul, M. Gales, and K. Knill, "Can generative large language models perform asr error correction?"
[4] S. Radhakrishnan et al., "Whispering llama: A cross-modal generative error correction framework for speech recognition"

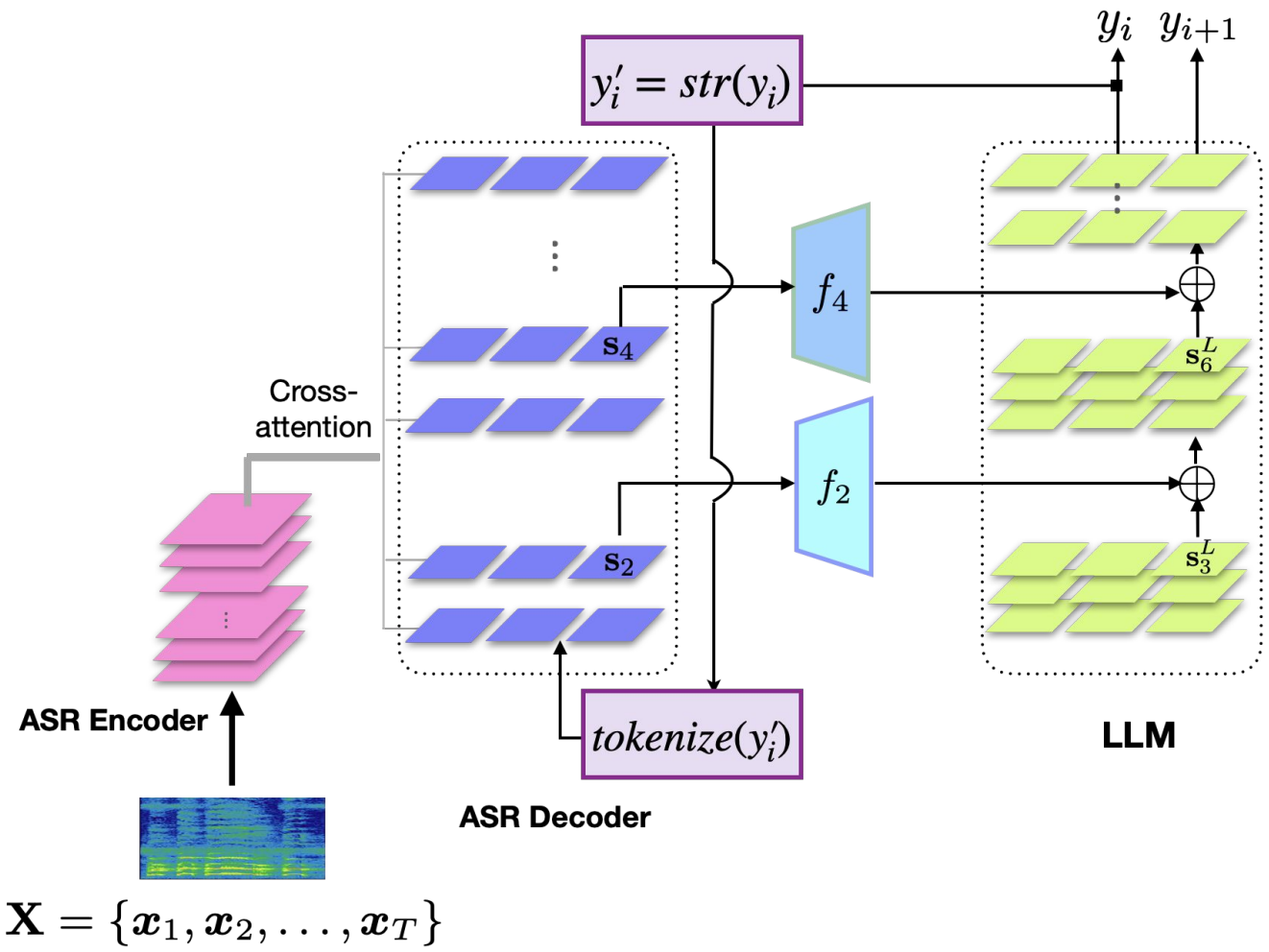
SALSA (Speedy ASR-LLM Synchronous Aggregation)



ASR-LLM coupling:

- The LLM generates the next token autoregressively.
- Connect ASR decoder state(s) via projection layers to the LLM.

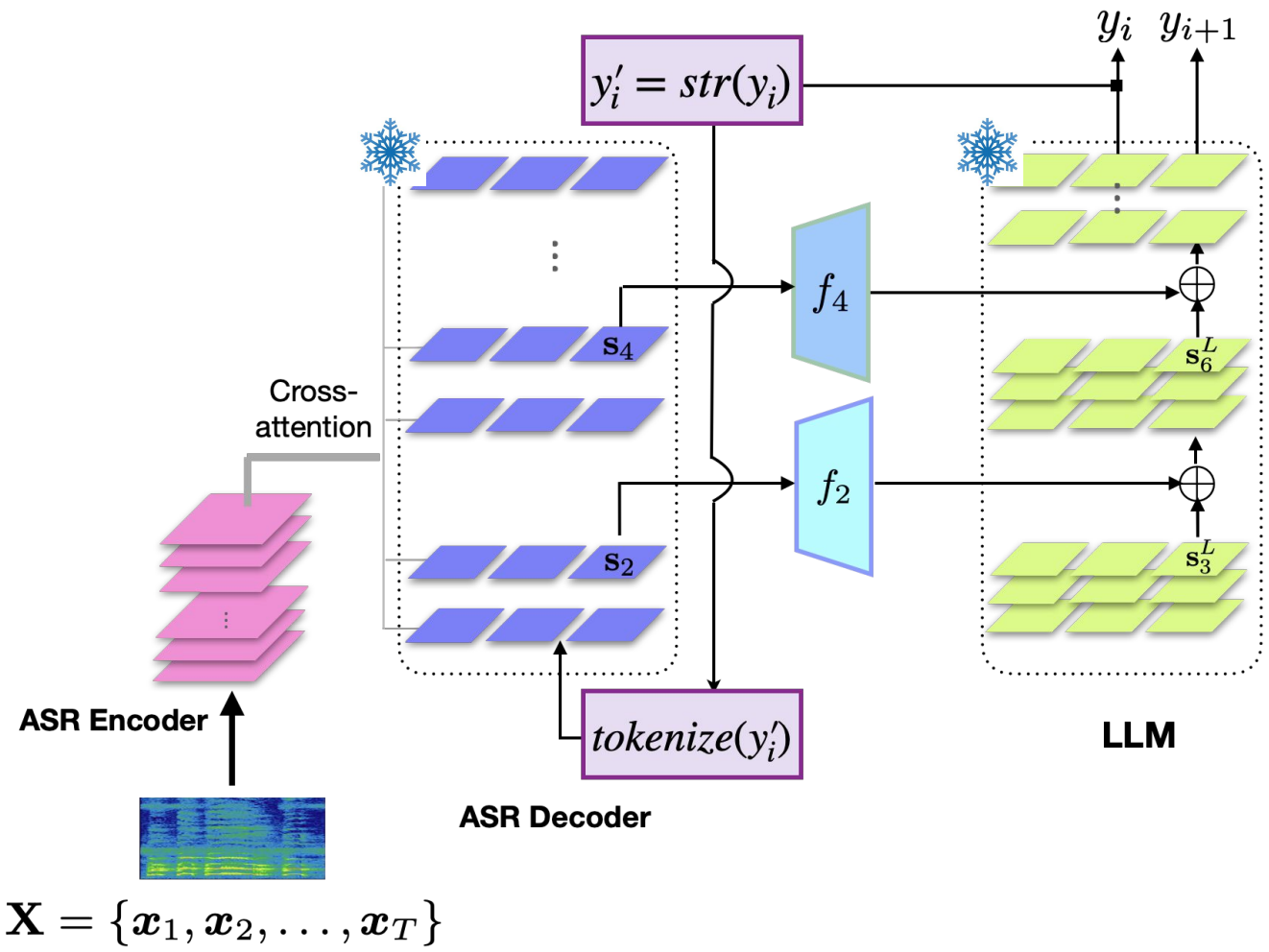
SALSA (Speedy ASR-LLM Synchronous Aggregation)



ASR-LLM coupling:

- The LLM generates the next token autoregressively.
- Connect ASR decoder state(s) via projection layers to the LLM.
- LLM predictions are fed back to ASR decoder. (Synchronous)

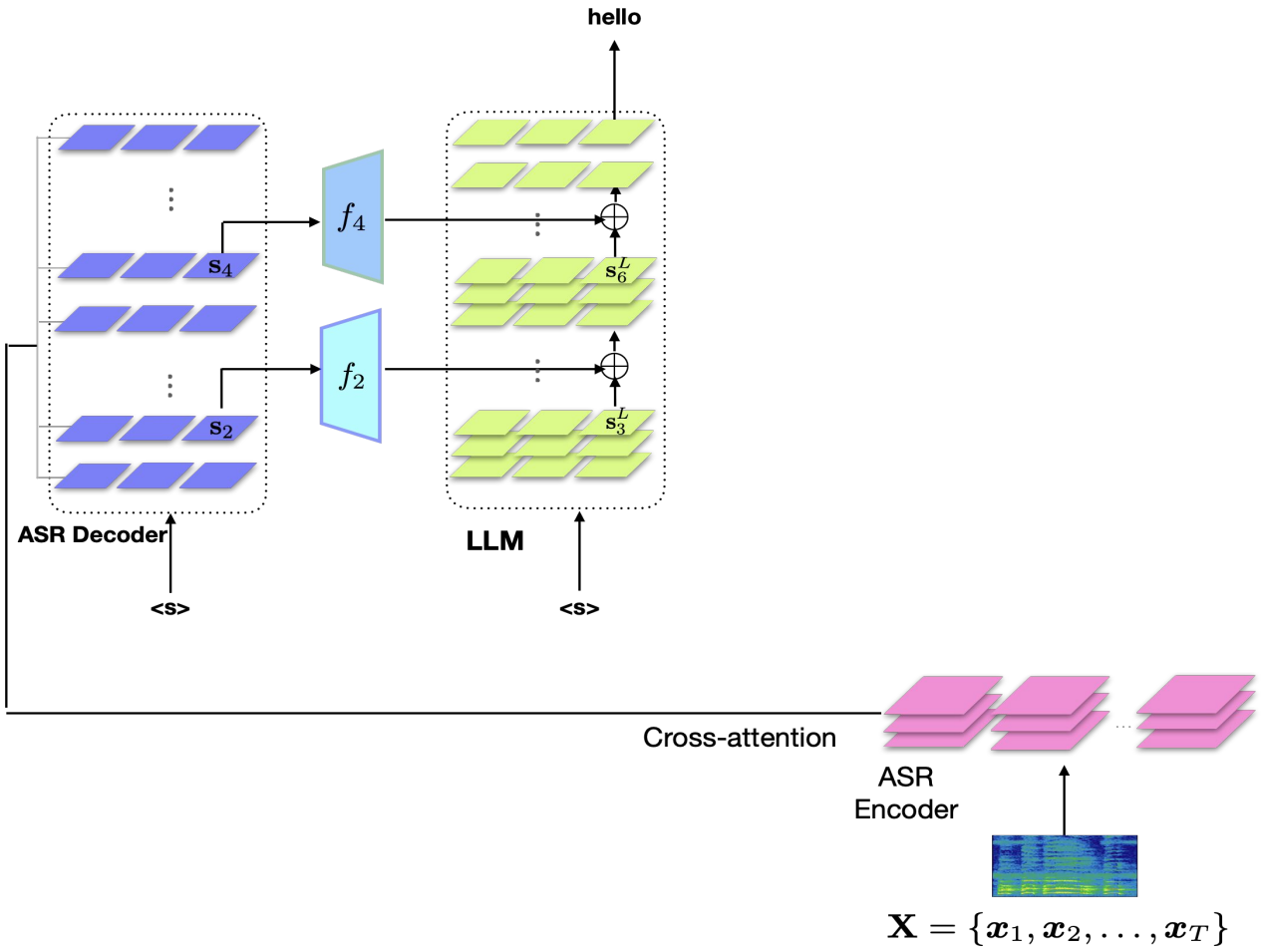
SALSA (Speedy ASR-LLM Synchronous Aggregation)



ASR-LLM coupling:

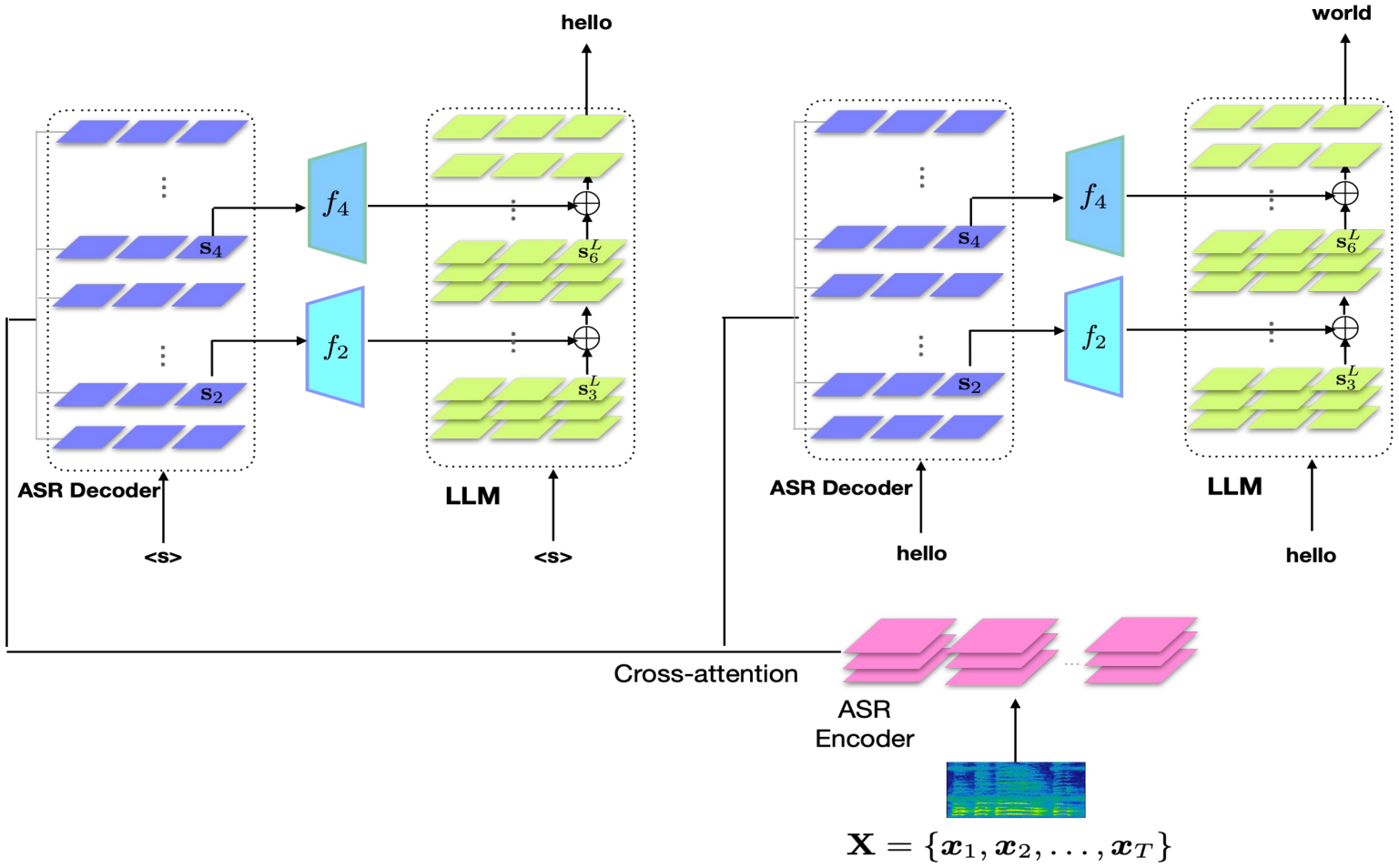
- The LLM generates the next token autoregressively.
- Connect ASR decoder state(s) via projection layers to the LLM.
- LLM predictions are fed back to ASR decoder. (Synchronous)
- ASR and LLM are frozen.

SALSA (Working Example)



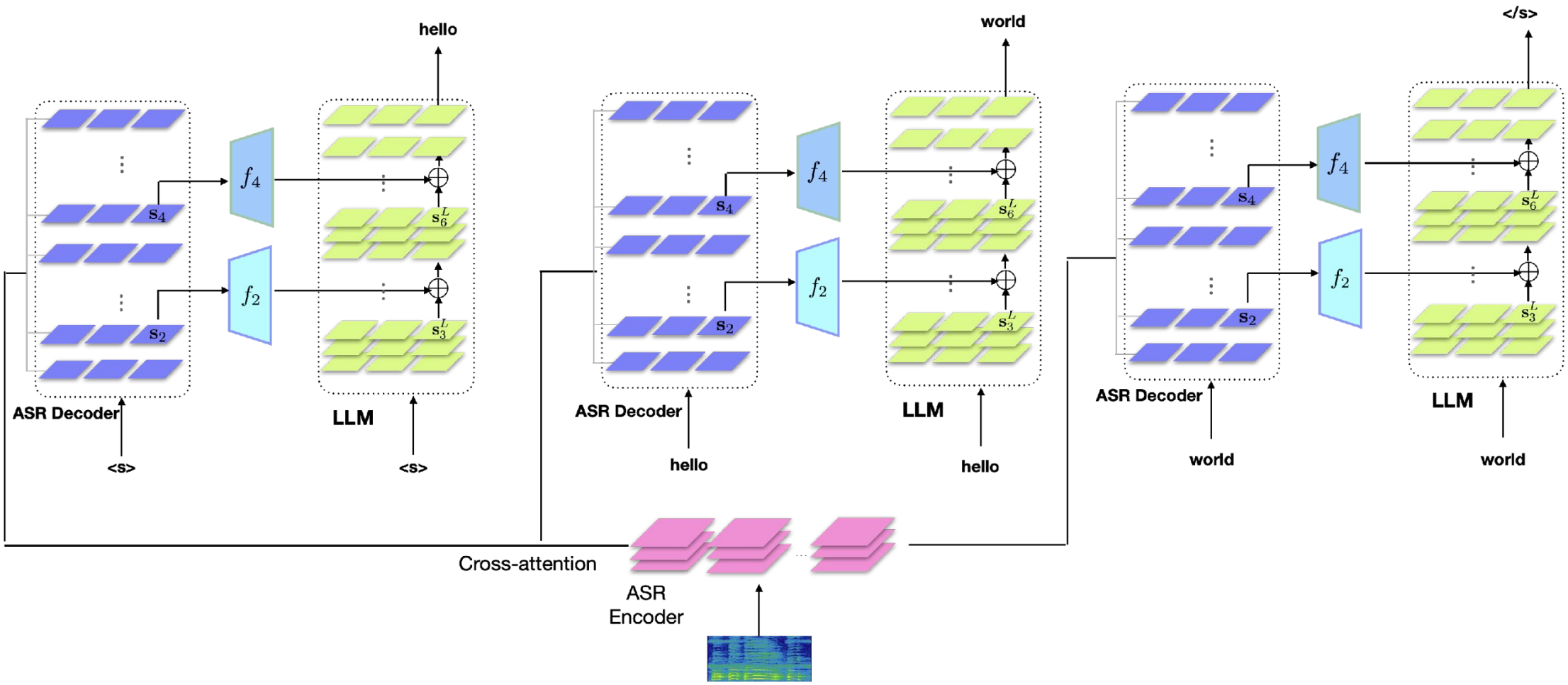
Ground Truth: hello world

SALSA (Working Example)



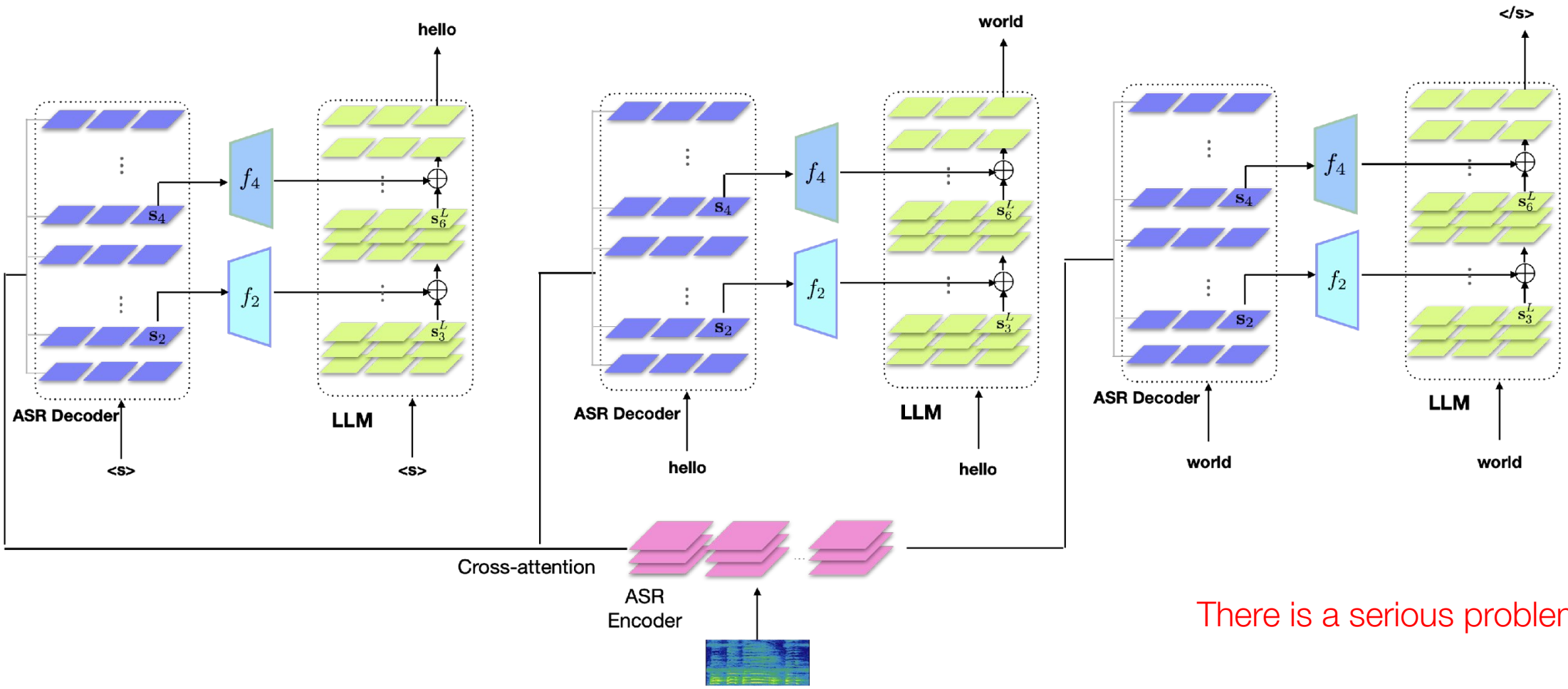
Ground Truth: hello world

SALSA (Working Example)



Ground Truth: hello world

SALSA (Working Example)

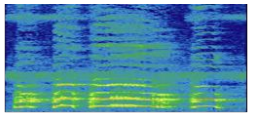


Ground Truth: hello world

There is a serious problem?

SALSA (Working Example)

Most likely ASR and LLM have different vocabularies and tokenizations.



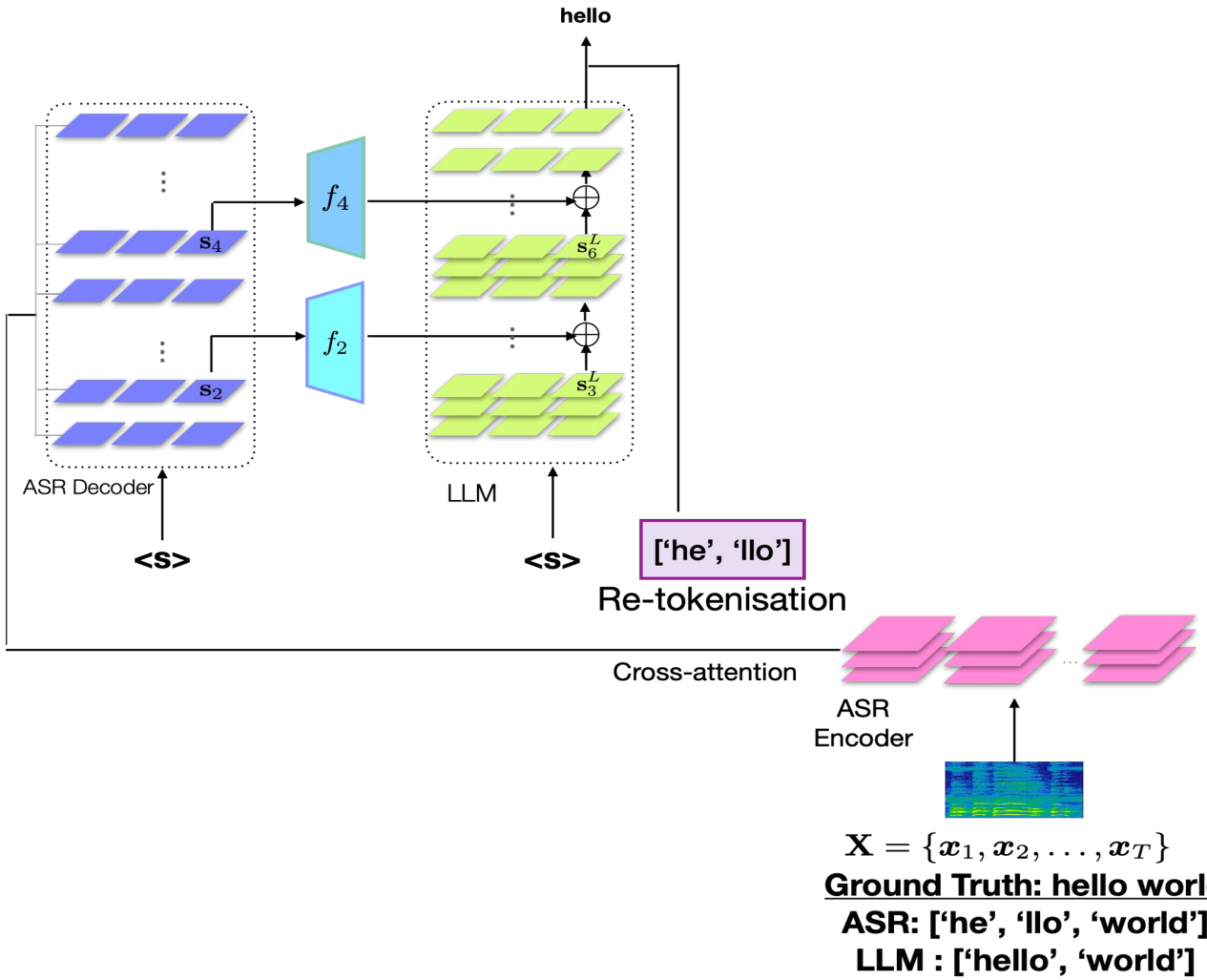
$$\mathbf{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_T\}$$

Ground Truth: hello world

Whisper : [‘he’, ‘llo’, ‘world’]

Llama-2 : [‘hello’, ‘world’]

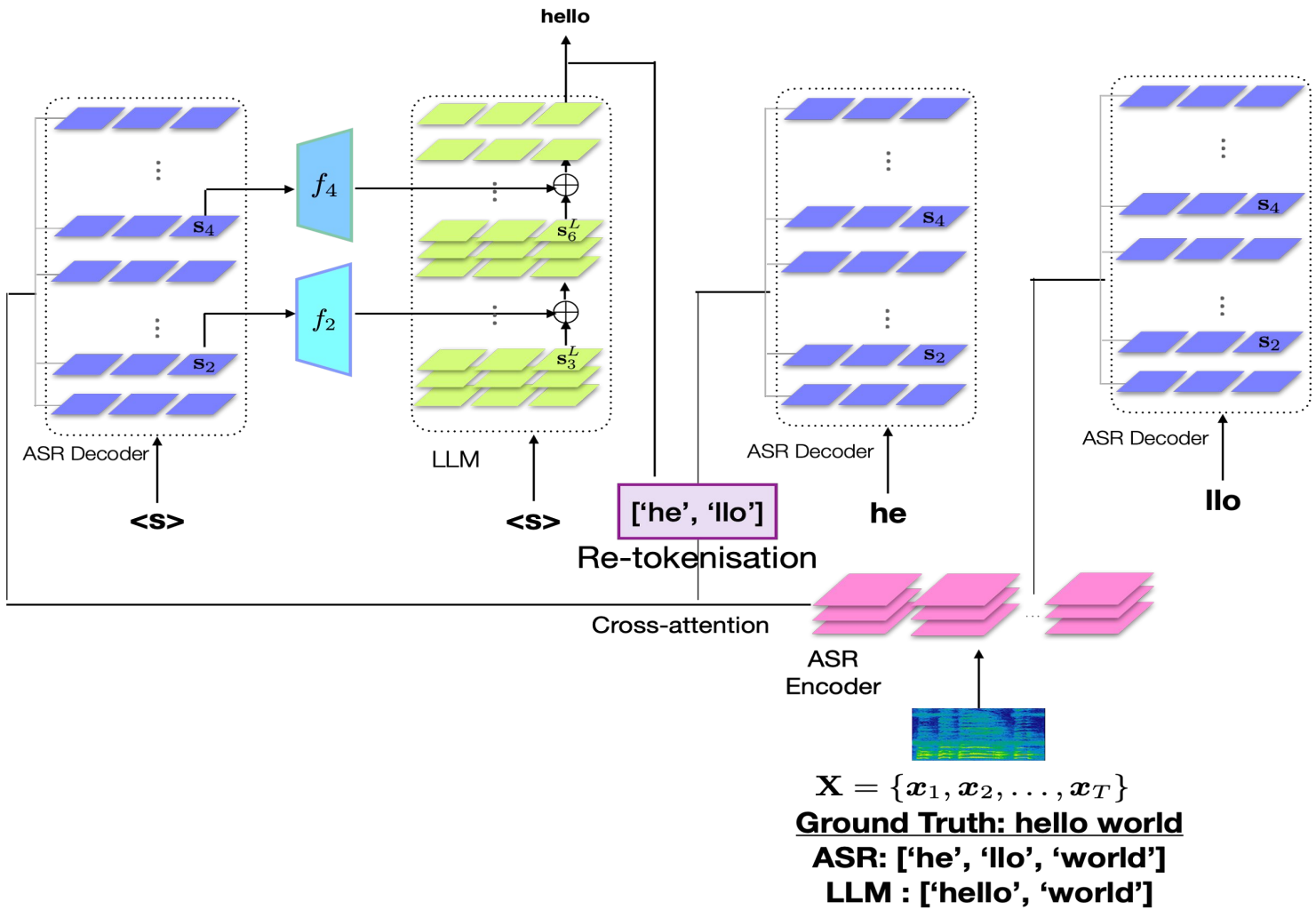
SALSA (Working Example)



ASR-LLM coupling:

- LLM output is re-tokenized with ASR tokenizer

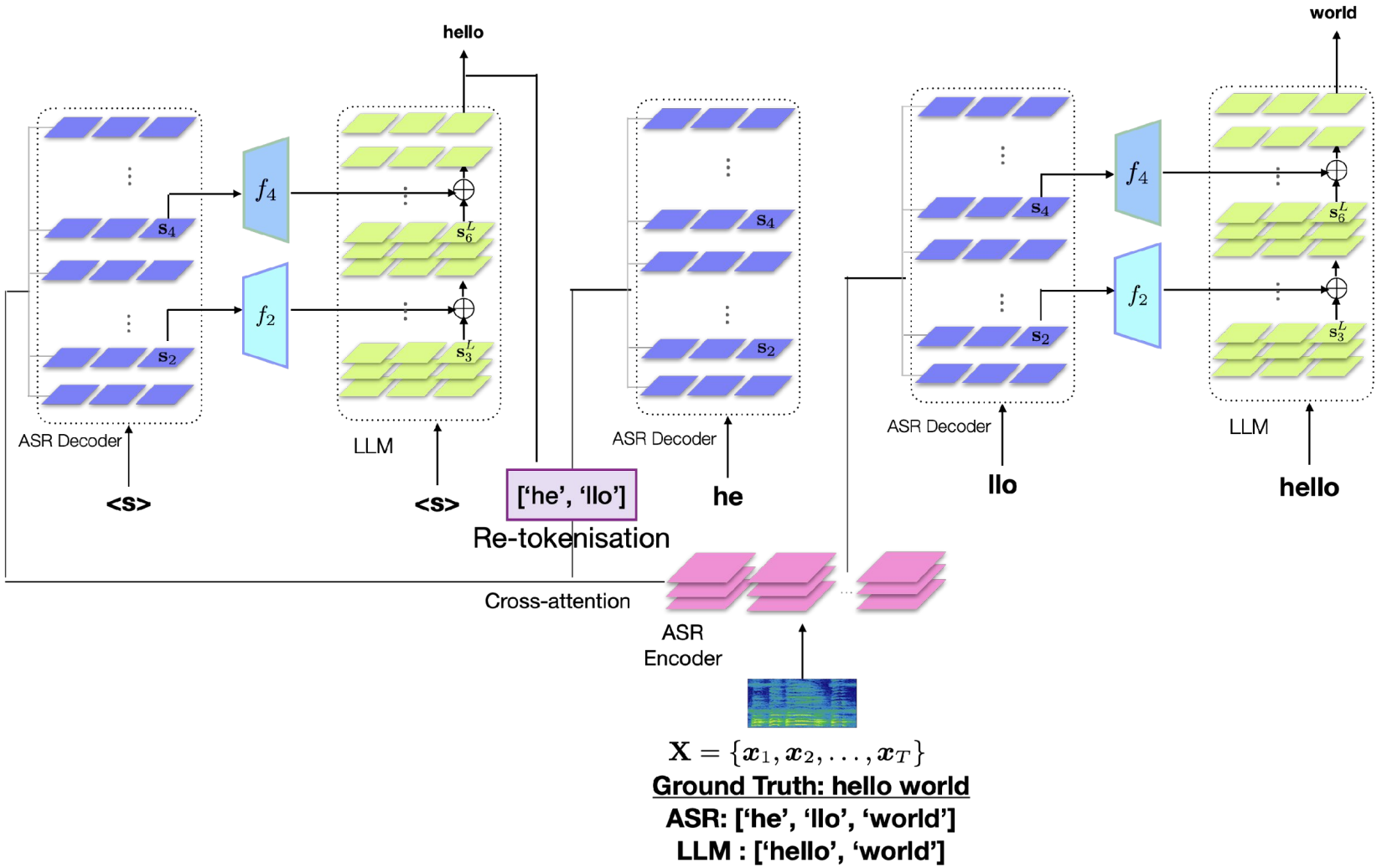
SALSA (Working Example)



ASR-LLM coupling:

- LLM output is re-tokenized with ASR tokenizer
- ASR Decoder is advanced with re-tokenized tokens

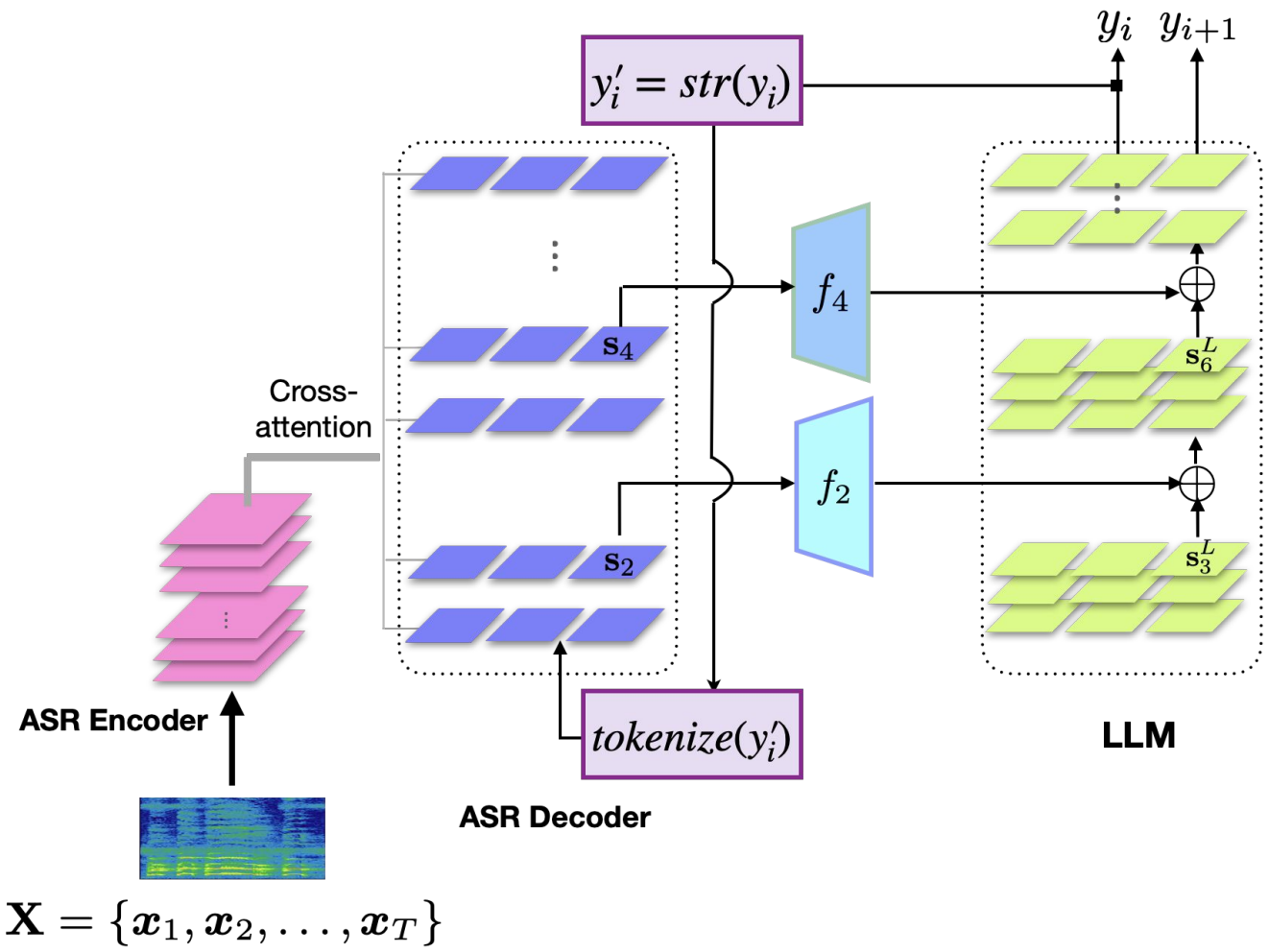
SALSA (Working Example)



ASR-LLM coupling:

- LLM output is re-tokenized with ASR tokenizer
- ASR Decoder is advanced with re-tokenized tokens
- Most recent ASR decoder state is then aggregated by LLM

SALSA (Speedy ASR-LLM Synchronous Aggregation)



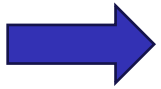
Training

- Tokenize the gold transcript using the LLM tokenizer, followed by the tokenization via the ASR model.
- Standard cross-entropy based training using probabilities from the LLM-decoder.
- Coupling strategy utilizing only the last decoder states is sufficient since both ASR/LLM decoders are autoregressive.

SALSA Usefulness

Limitations:

- N-best lists unreliable for low resource languages



SALSA Fixes:

- Synchronous integration of ASR decoder with LLM

SALSA Usefulness

Limitations:

- N-best lists unreliable for low resource languages
- High Latency

SALSA Fixes:

- Synchronous integration of ASR decoder with LLM
- Offers one pass integration

SALSA Usefulness

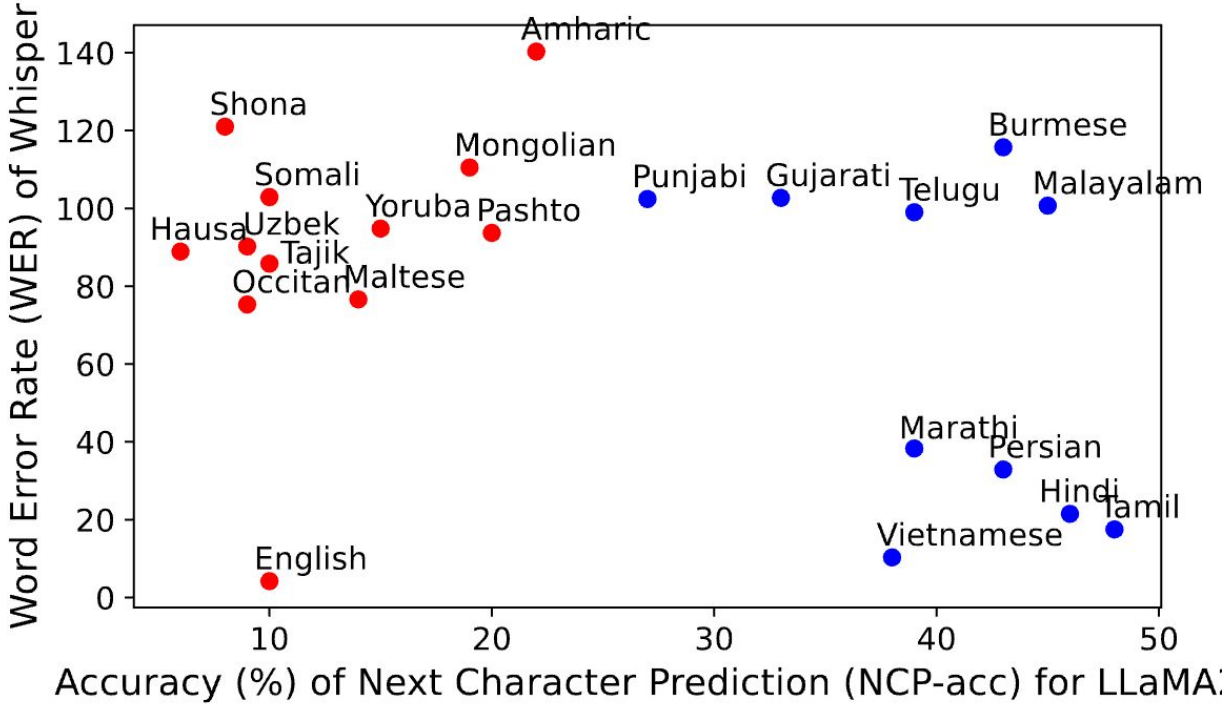
Limitations:

- N-best lists unreliable for low resource languages
- High Latency
- Large training data to learn cross-attention

SALSA Fixes:

- Synchronous integration of ASR decoder with LLM
- Offers one pass integration
- Leverages rich ASR decoder states

Optimal conditions for SALSA



Language selection for SALSA:

- Next Character Prediction (NCP-acc) vs. Word Error Rate (WER %)
- We chose 8 languages that have high NCP-acc and medium to high WER using Whisper.

Results on Eight Low-resource Languages from FLEURS

Method	# params	Gujarati	Hindi	Malayalam	Marathi	Persian	Punjabi	Tamil	Telugu	Average
Our Baselines (Reproduced using official repositories)										
Whisper (Large-v2) [10]	–	108.2	35.9	107.8	84.7	35.8	101.8	48.0	104.4	78.3
w/ LoRA fine-tuning [35]	15M	**55.7**	**19.3**	**54.9**	**35.8**	**16.9** †	**46.7**	**38.0** †	**47.5**	**39.4**
Whispering-LlaMA [6]	26M	90.8	53.2	101.1	105.9	86.1	92.6	89.1	106.4	90.7
SALSA-7B (w/ LLaMA2-7B)										
w/ Whisper (Large-v2)	17M	37.8	18.2	40.9	37.5	18.6	37.0	40.1	44.9	34.4
w/ LoRA fine-tuned Whisper	17M	**34.6** †	17.6	35.3	**35.6** †	18.2	34.8	42.6	45.1	33.0
SALSA-13B (w/ LLaMA2-13B)										
w/ Whisper (Large-v2)	19M	37.1	17.4	40.4	37.8	18.6	37.0	40.1	45.0	34.1
w/ LoRA fine-tuned Whisper	19M	34.9	**16.8** †	**34.8** †	36.5	17.6	**34.5** †	41.4	**45.0** †	**32.7** †

- ASR: Whisper Large-v2
- LLM: Llama2-7B, Llama-13B
- Data: 12 hours of labeled speech in each language

- Large WER reductions for some languages (e.g., Gujarati, Malayalam).
- Whispering-Llama suffers due to poor quality of n-best lists for low-resource languages.

Results of multilingual SALSA

Method	# params	Gujarati	Hindi	Malayalam	Marathi	Punjabi	Tamil	Telugu	Average
Whisper (Large-v2) [10]	–	108.2	35.9	107.8	84.7	101.8	48.0	104.4	84.4
w/ LoRA fine-tuning [35]	15M	56.8	19.8	**39.8** †	37.4	55.8	**39.8** †	50.5	42.8
SALSA-7B w/ Whisper (Large-v2)	17M	**36.4** †	**17.5** †	40.0	**36.7** †	**36.1** †	40.2	**43.2** †	**35.8** †

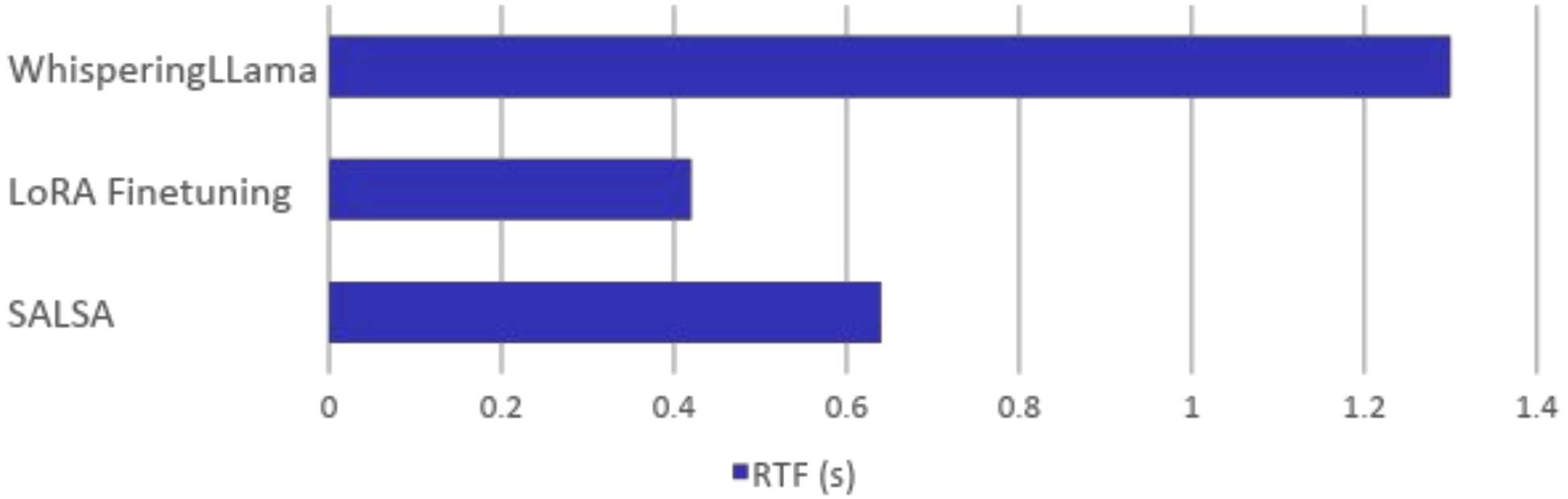
- ASR: Whisper Large-v2
- LLM: Llama2-7B, Llama-13B
- Data: Single SALSA model trained on 7 Indic languages.

Multilingual SALSA outperforms the LORA fine-tuned Whisper by **20%**

Training Complexity of SALSA

Training Time for 12 hours of audio from Fleurs dataset

Inference Complexity of SALSA



Comparison of Real Time Factor (RTF) for different methods.

Decoder Coupling

Method	# params	Average WER
SALSA-7B (w/ LLaMA2-7B)		
8 adapter layers (uniformly distributed)	17M	34.4
8 adapter layers (all at the end)	17M	34.8
4 adapter layers (uniformly distributed)	8.5M	36.4
4 adapter layers (all at the end)	8.5M	36.9

Comparing the number of adapter layers and the position of adapter layers averaged across 8 languages.

- Reducing adaptation layers in SALSA-7B and SALSA-13B causes a 5% WER increase but still outperforms LoRA finetuning, which averages a 39.4% WER.
- Adaptation layers uniformly distributed outperforms the adapter layers at the end.

Conclusion

- SALSA: Combines ASR system expertise with LLM's superior language modelling.
- WER Reduction: Achieves significant WER improvements over fine-tuning the ASR model alone.
- Efficiency: Offers one-pass decoding and faster training compared to other LLM-ASR fusion methods.

Code: https://github.com/csalt-research/salsa
Arxiv: http://arxiv.org/abs/2408.16542

Thank You!

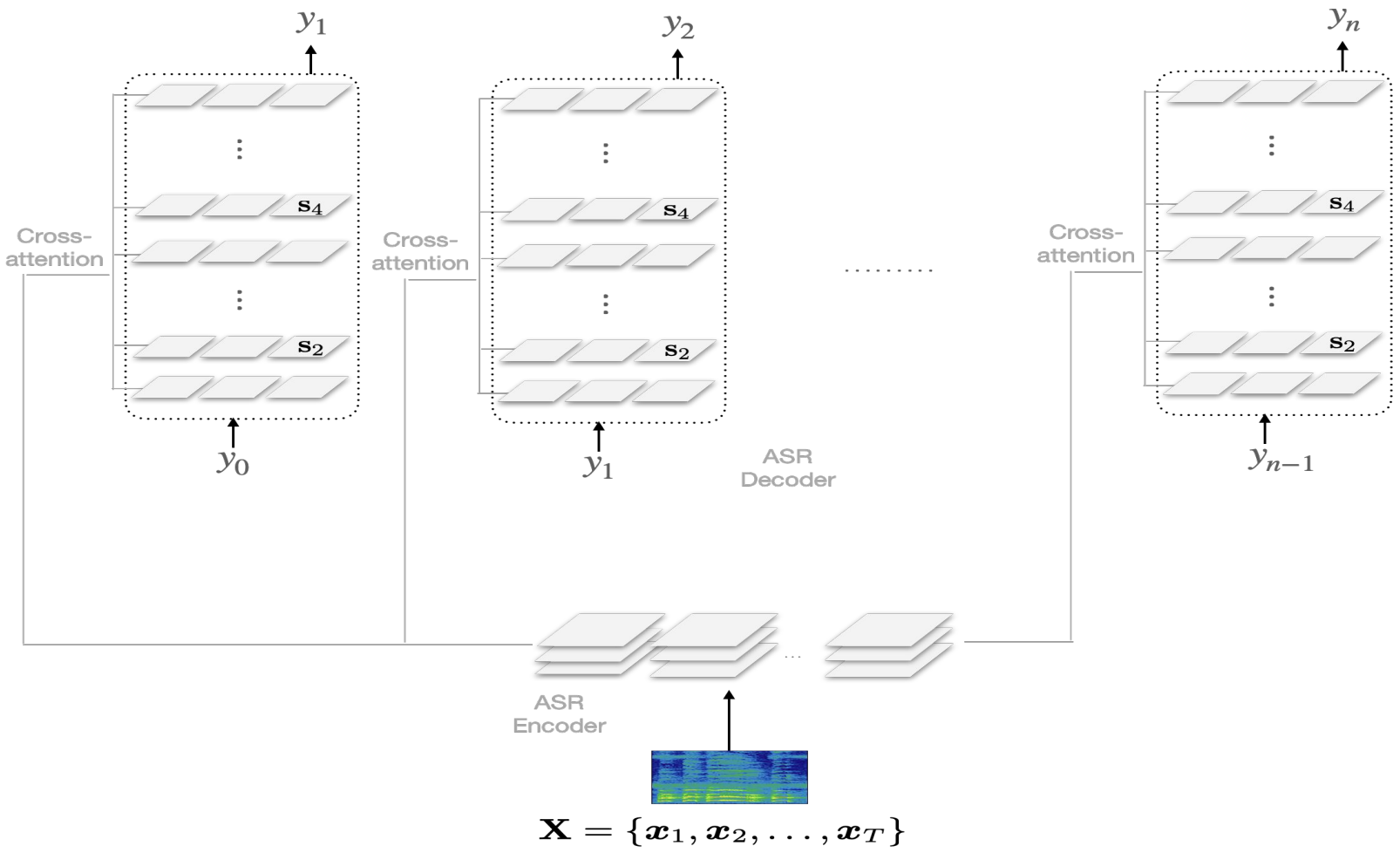
Backup

SALSA English Results

	WER	CER
Whisper (large-v2)	4.6	2.5
SALSA Llama2 7B (w/ Whisper large-v2)	2.6	0.9

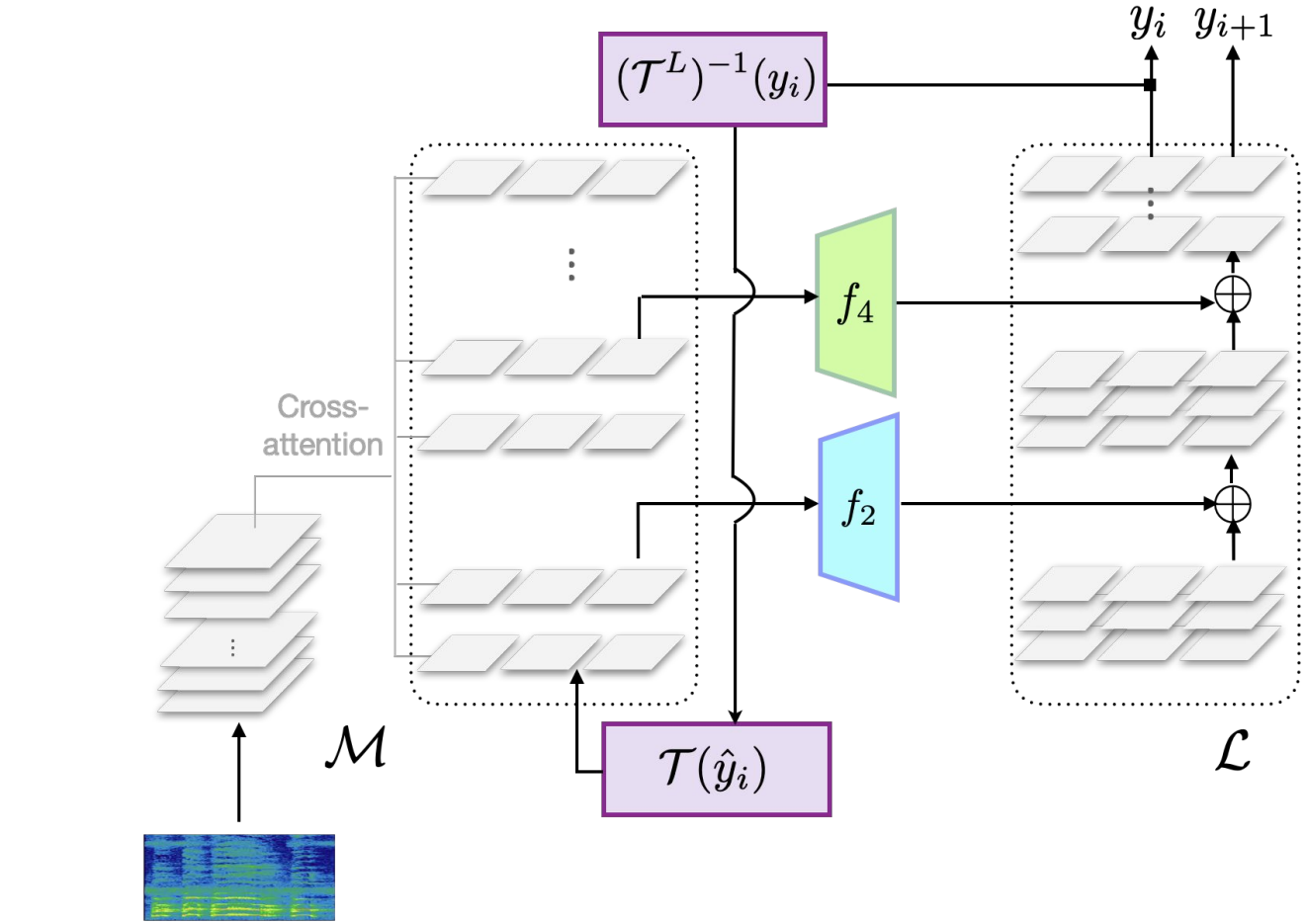
- Results on test-clean.
- SALSA trained on Librispeech 960 hours.

ASR Encoder-Decoder Architecture



- Autoregressive decoding.
- Cross-attention to encoder features.

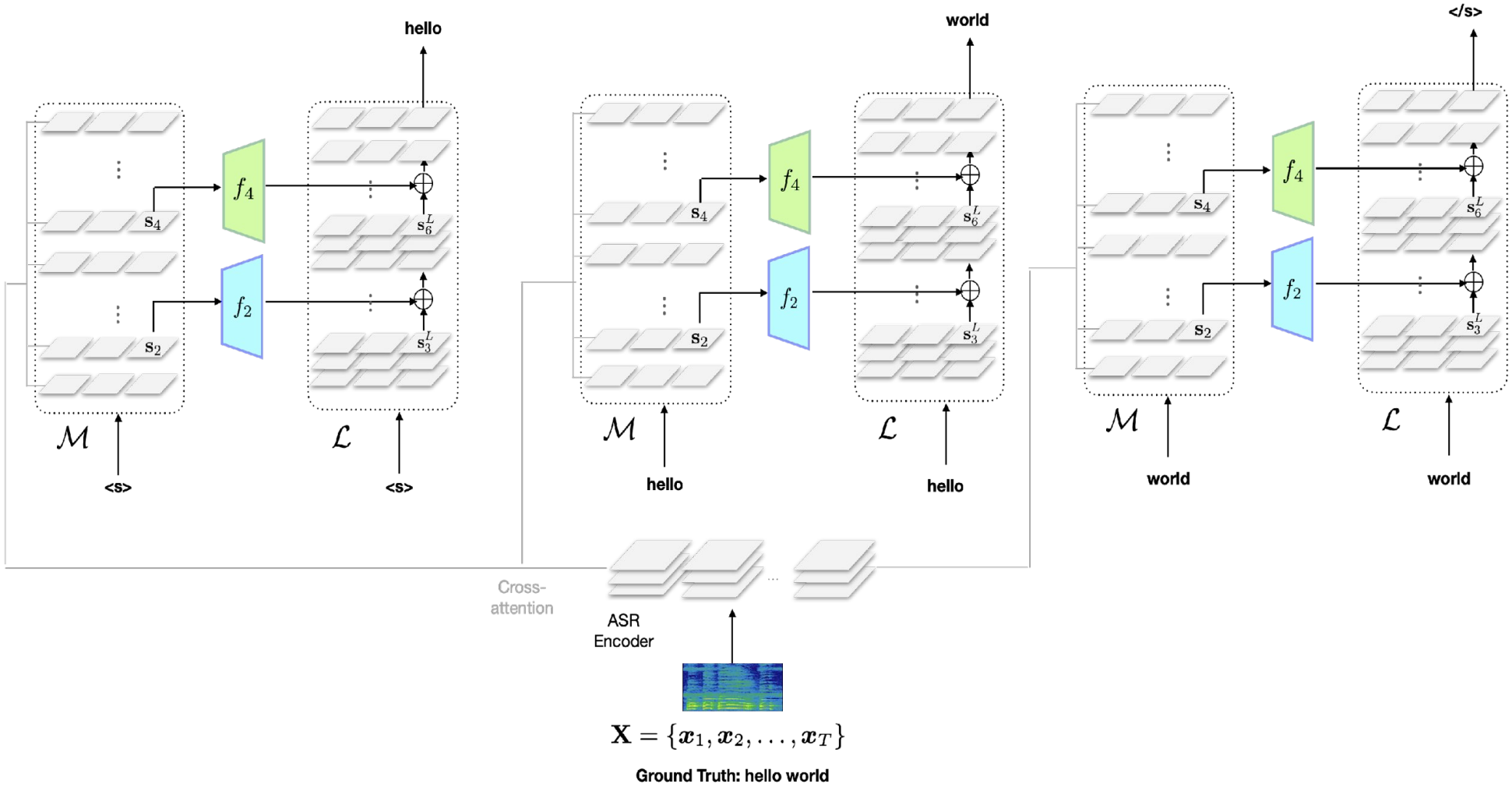
SALSA (Speedy ASR-LLM Synchronous Aggregation)



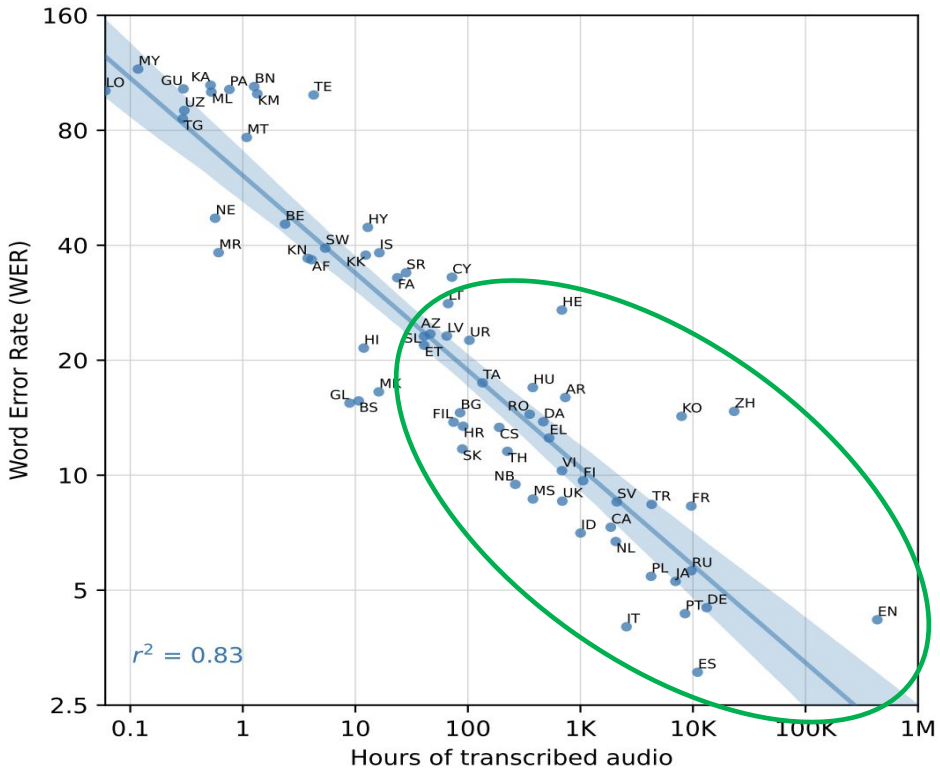
Generating the Output:

- SALSA can work with different ASR/LLM vocabularies and tokenizers.
- The LLM generated text is re-tokenized with the ASR tokenizer and the decoder state of the ASR model is advanced with the newly predicted tokens.
- The updated decoder state is then used by the LLM in its subsequent iterations.

SALSA (Speedy ASR-LLM Synchronous Aggregation)



ASR performance

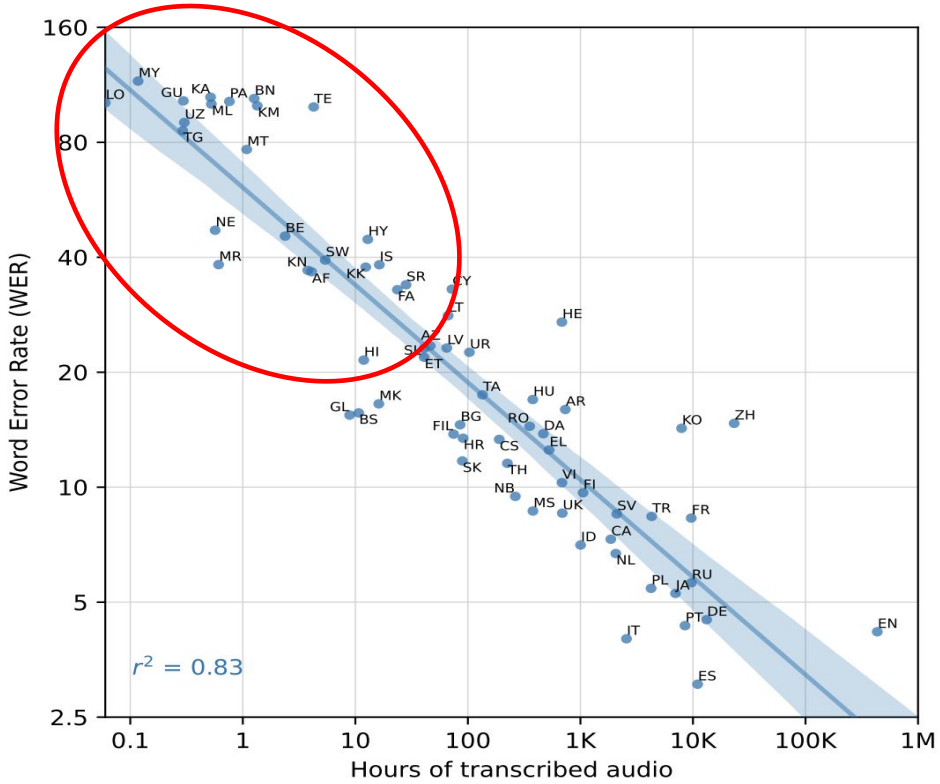


High-Resource Languages:

- Abundant transcribed audio enhances performance.

Reference: Radford, Alec, et al. "Robust speech recognition via large-scale weak supervision.

ASR performance



High-Resource Languages:

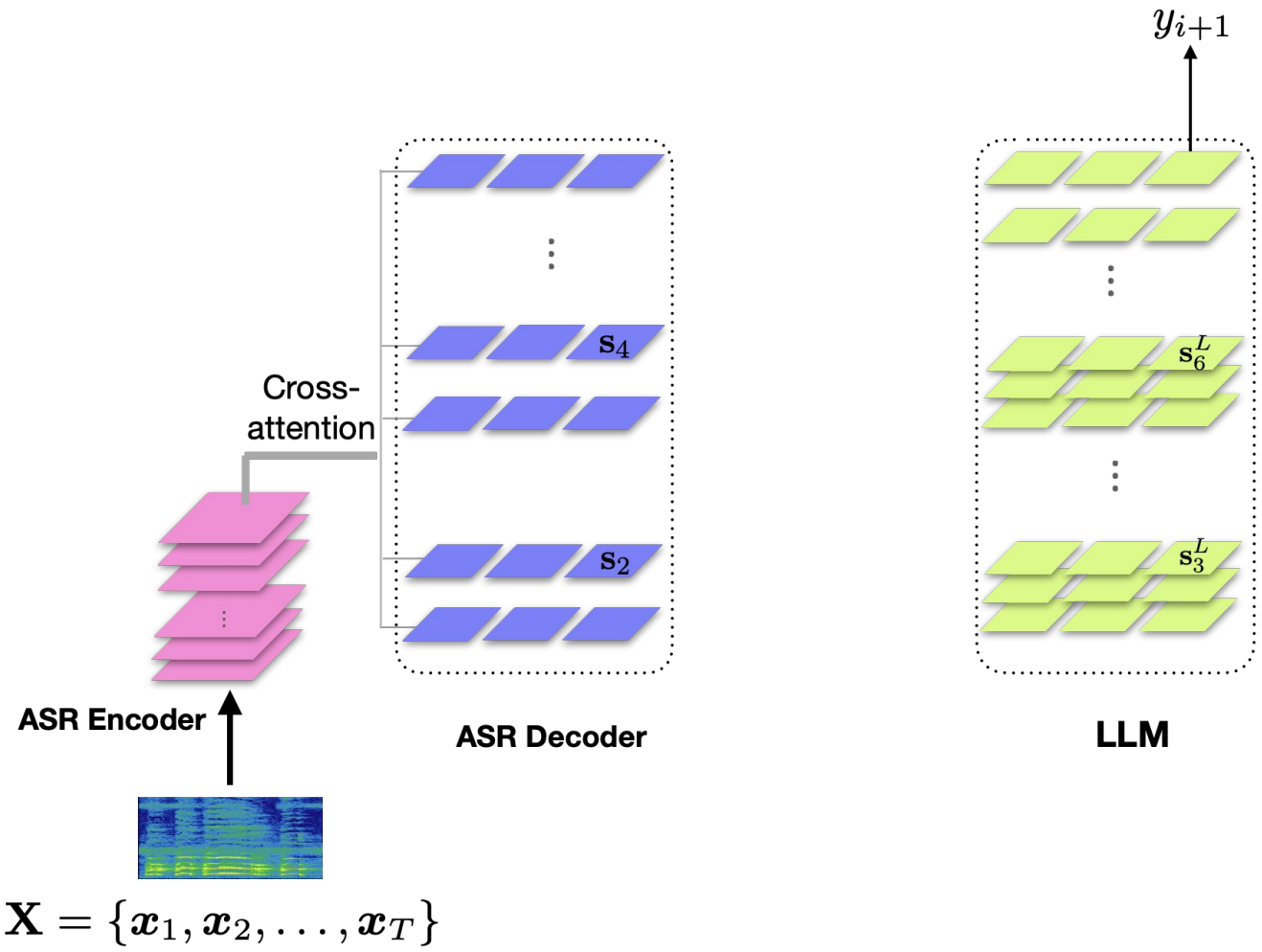
- Abundant transcribed audio enhances performance.

Low-Resource Languages:

- Performance remains suboptimal due to limited data availability.
- Acquiring transcribed audio is particularly challenging.

Reference: Radford, Alec, et al. "Robust speech recognition via large-scale weak supervision.

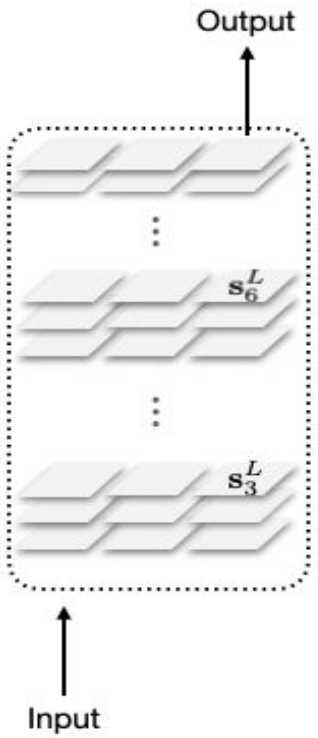
SALSA (Speedy ASR-LLM Synchronous Aggregation)



ASR-LLM coupling:

- The LLM generates the next token autoregressively.

Large Language Models



Large Language Model

- **Massive Scale:** Trained on huge datasets enabling deep understanding of language.
- **Emergent Abilities:** Exhibit unexpected skills like translation and summarization without specific training.
- **Multilingual Understanding:** Understand and generate text across multiple languages.