



Data-Centric Lessons To Improve Speech-Language Pretraining

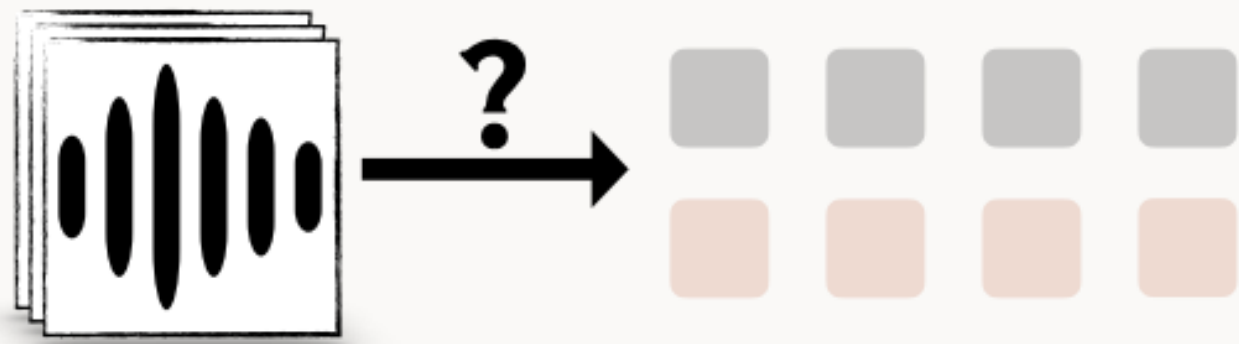
Vishaal Udandarao, Zhiyun Lu, Xuankai Chang, Yongqiang Wang, Violet Z. Yao, Albin M. Jose, Fartash Faghri, Josh Gardner, Chung-Cheng Chiu

ICLR'2026 | Apple

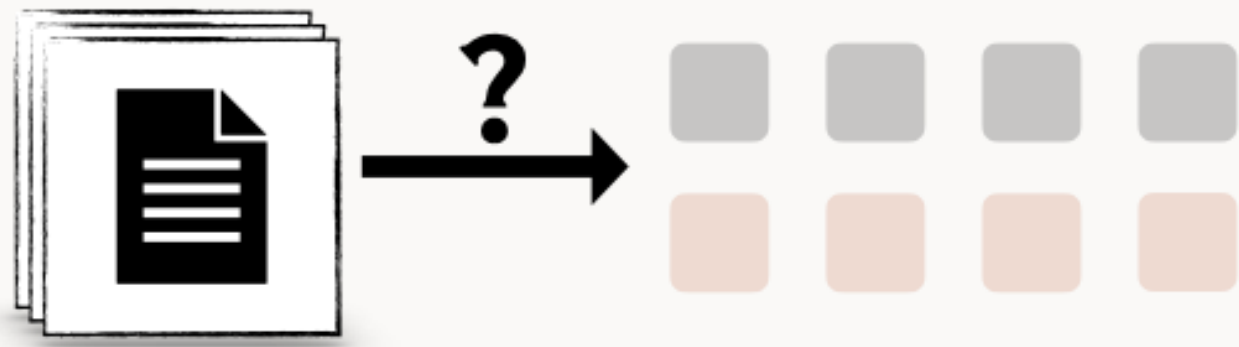
Overview

Our Three Data-Centric Research Questions

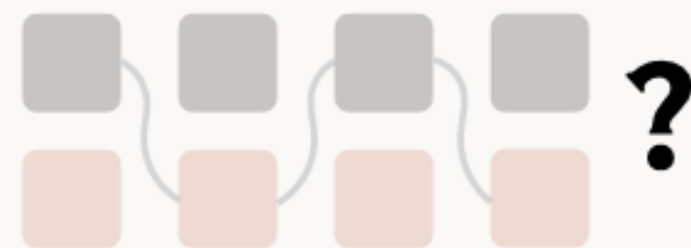
How to *process* raw audio into interleaved speech-text training data?



How to *construct* synthetic datasets using quality text-only datasets?

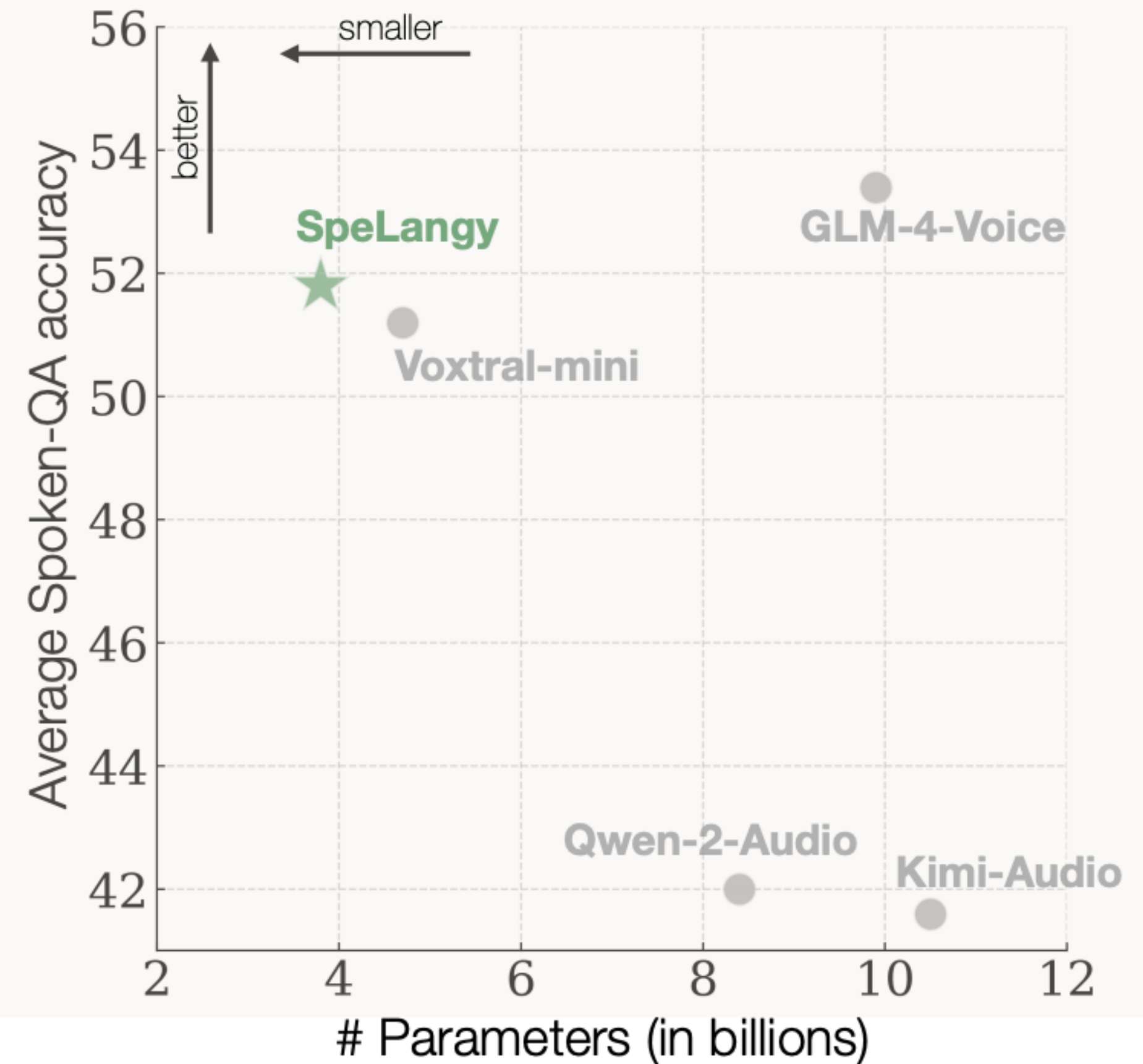


How to *interleave* between speech and text modalities while training?



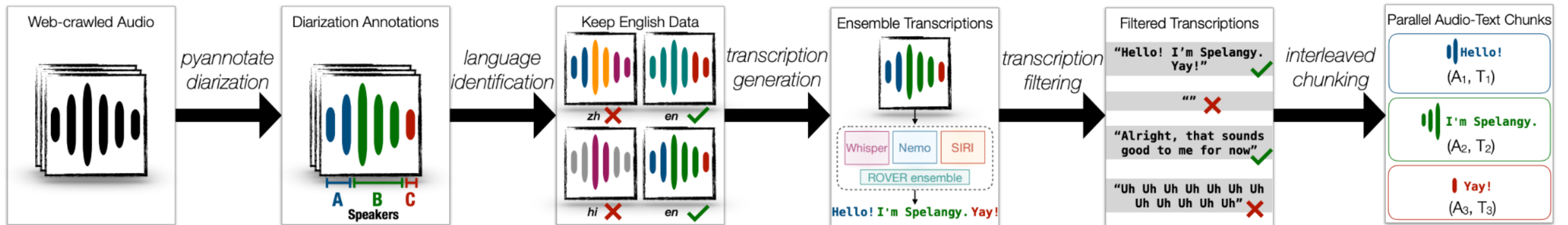
Speech token Text token

Our SpeLangy model is small and highly performant



Pipeline for processing web-crawled audio → speech-text training data

Processing web-crawled audio into speech-text interleaved data



RQ1: How to *process* raw audio into interleaved training data?

Sec 3.3 Coarse vs fine-grained interleaving

Diarized Audio with Paired Transcription



Hi. How are you? All good here. What's up?



Coarse-grained Interleaving



(A_1, T_1)

(A_2, T_2)

Fine-grained Interleaving



(A_1, T_1)

(A_2, T_2)

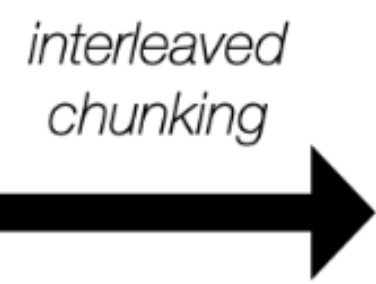
(A_3, T_3)

(A_4, T_4)

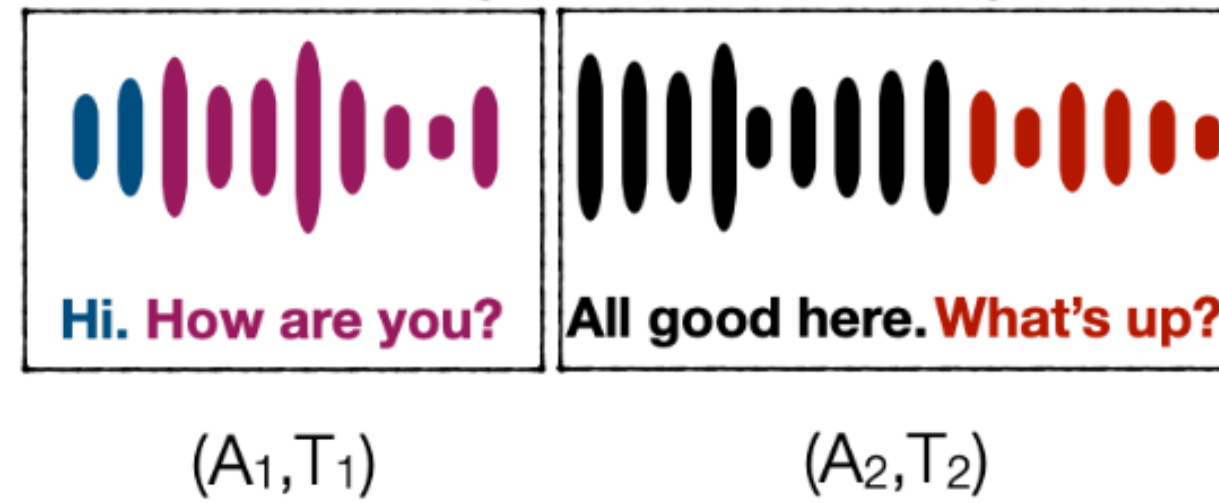
RQ1: How to *process* raw audio into interleaved training data?

Sec 3.3 Coarse vs fine-grained interleaving

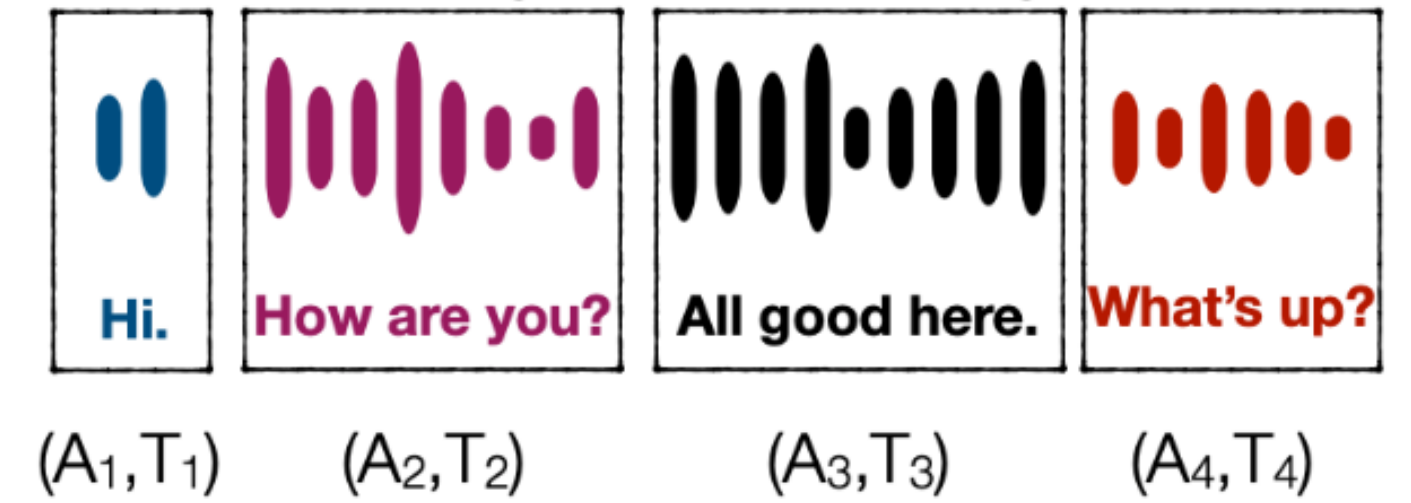
Diarized Audio with Paired Transcription



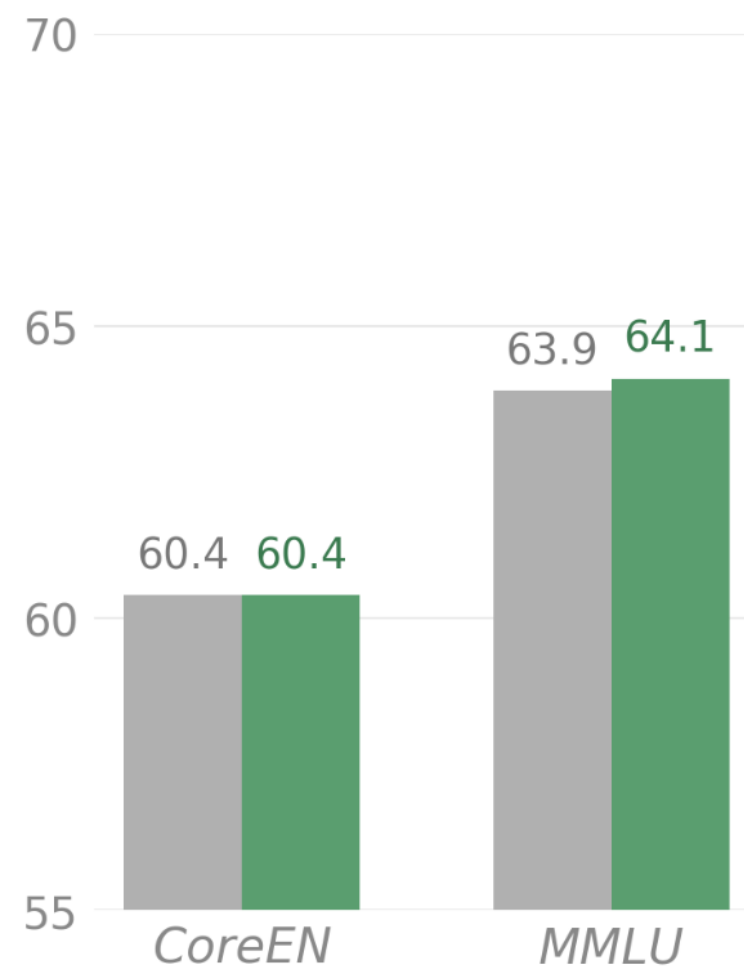
Coarse-grained Interleaving



Fine-grained Interleaving



Text understanding (T→T)

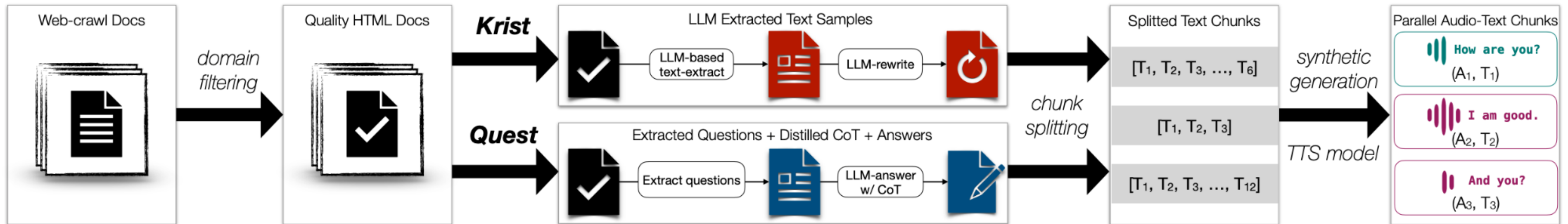


SQA accuracy (S→T) %



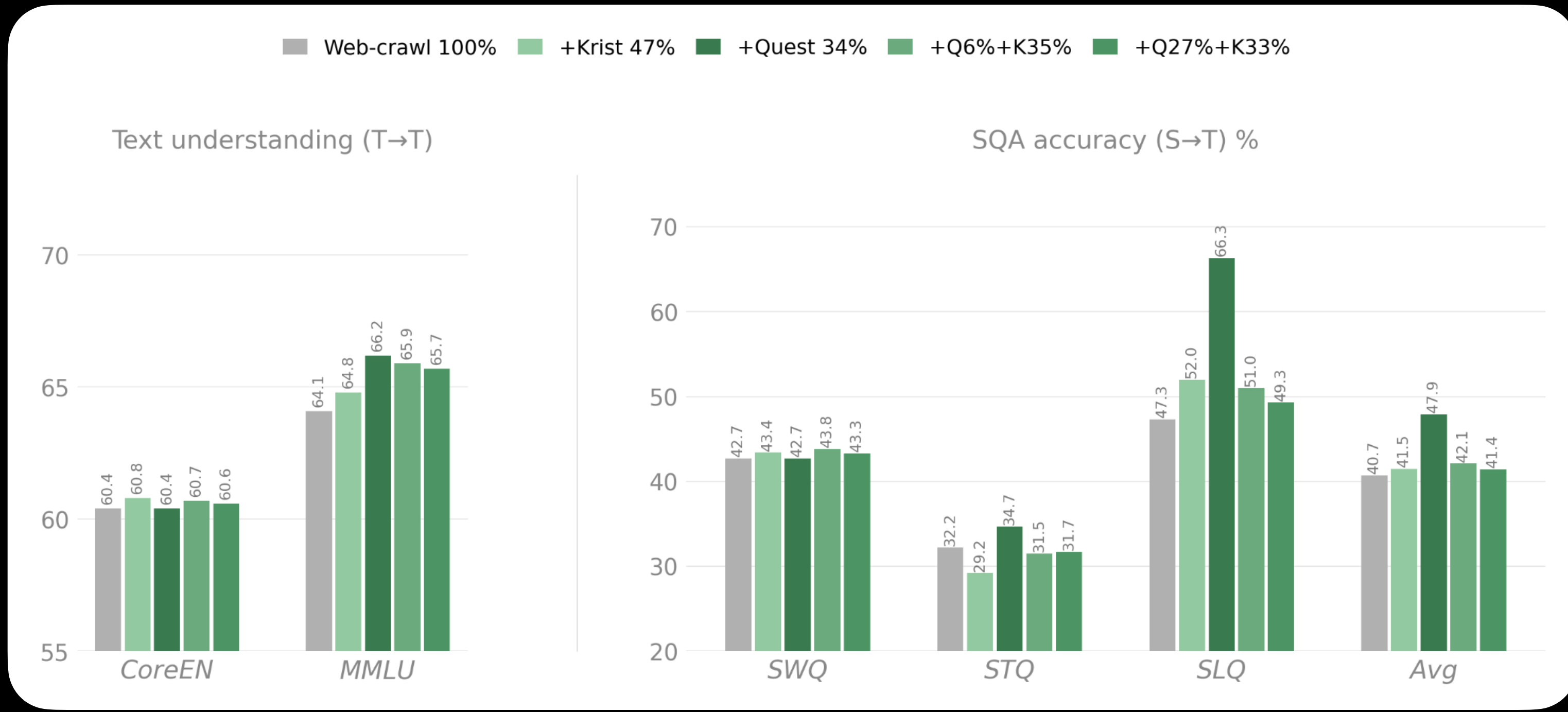
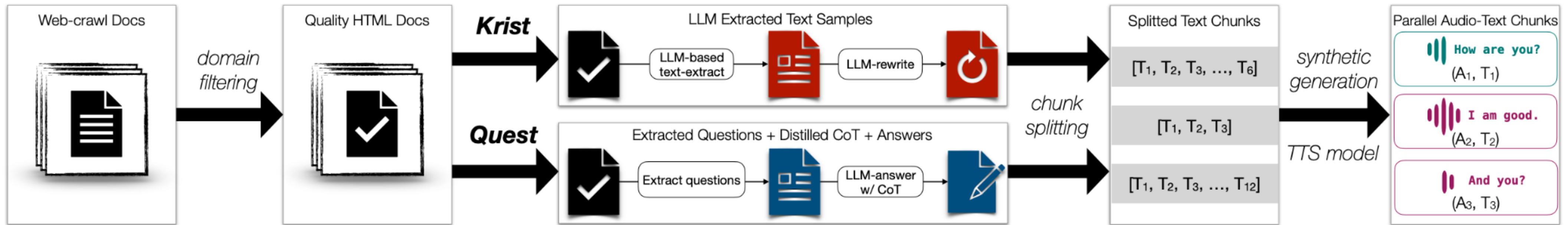
RQ2: How to *construct* synthetic speech-text data?

Sec 3.4 Generating synthetic speech-text interleaved data



RQ2: How to *construct* synthetic speech-text data?

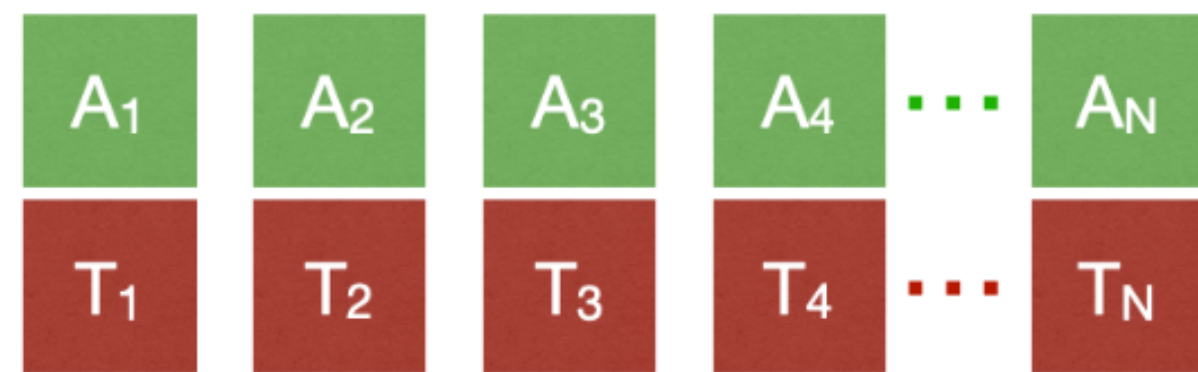
Sec 3.4 Generating synthetic speech-text interleaved data



RQ3: How to *interleave* between speech and text chunks?

Sec 3.5 Modality-sampling during interleaved training

Speech-Text Interleaved Training Sample



sampling during training



Deterministic Sampling



Switch modality every chunk for training

Stochastic Sampling

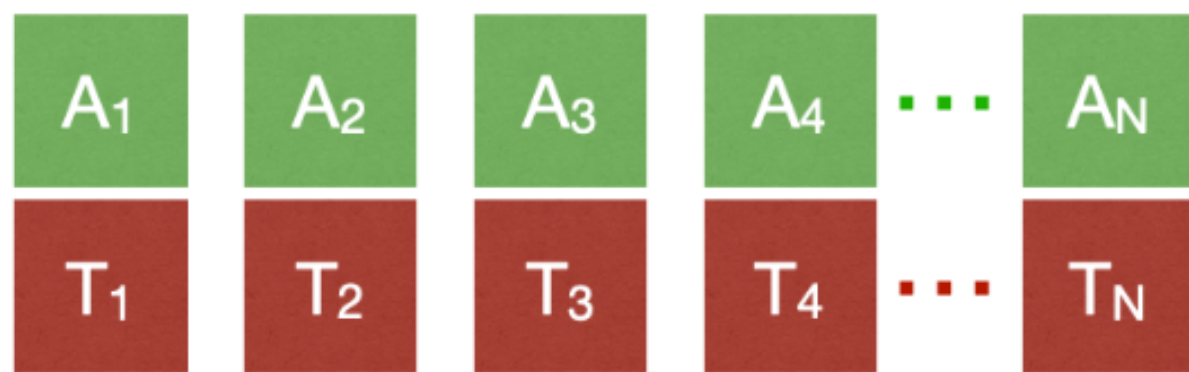


Randomly sample modality at each chunk

RQ3: How to *interleave* between speech and text chunks?

Sec 3.5 Modality-sampling during interleaved training

Speech-Text Interleaved Training Sample



Deterministic Sampling

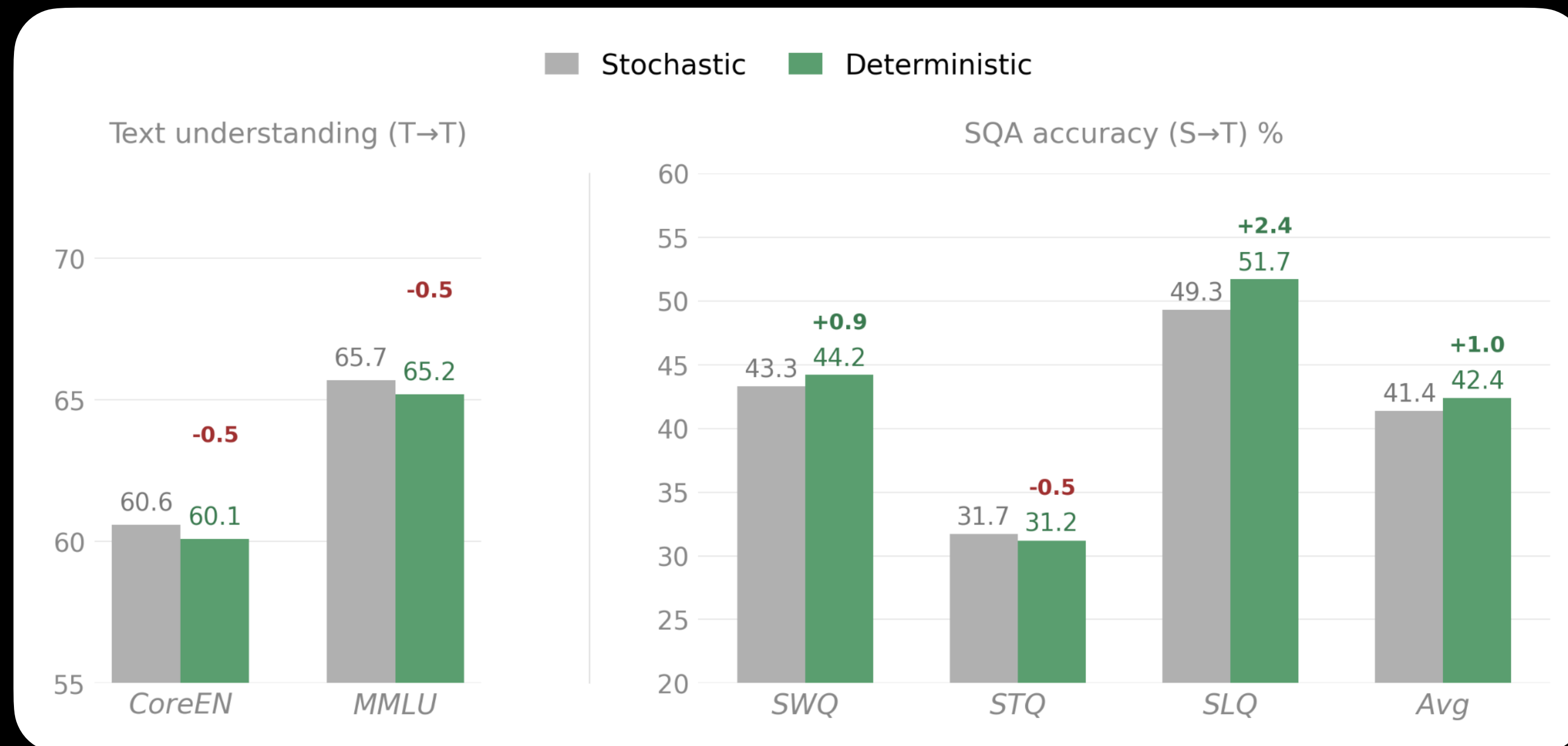


Switch modality every chunk for training

Stochastic Sampling



Randomly sample modality at each chunk



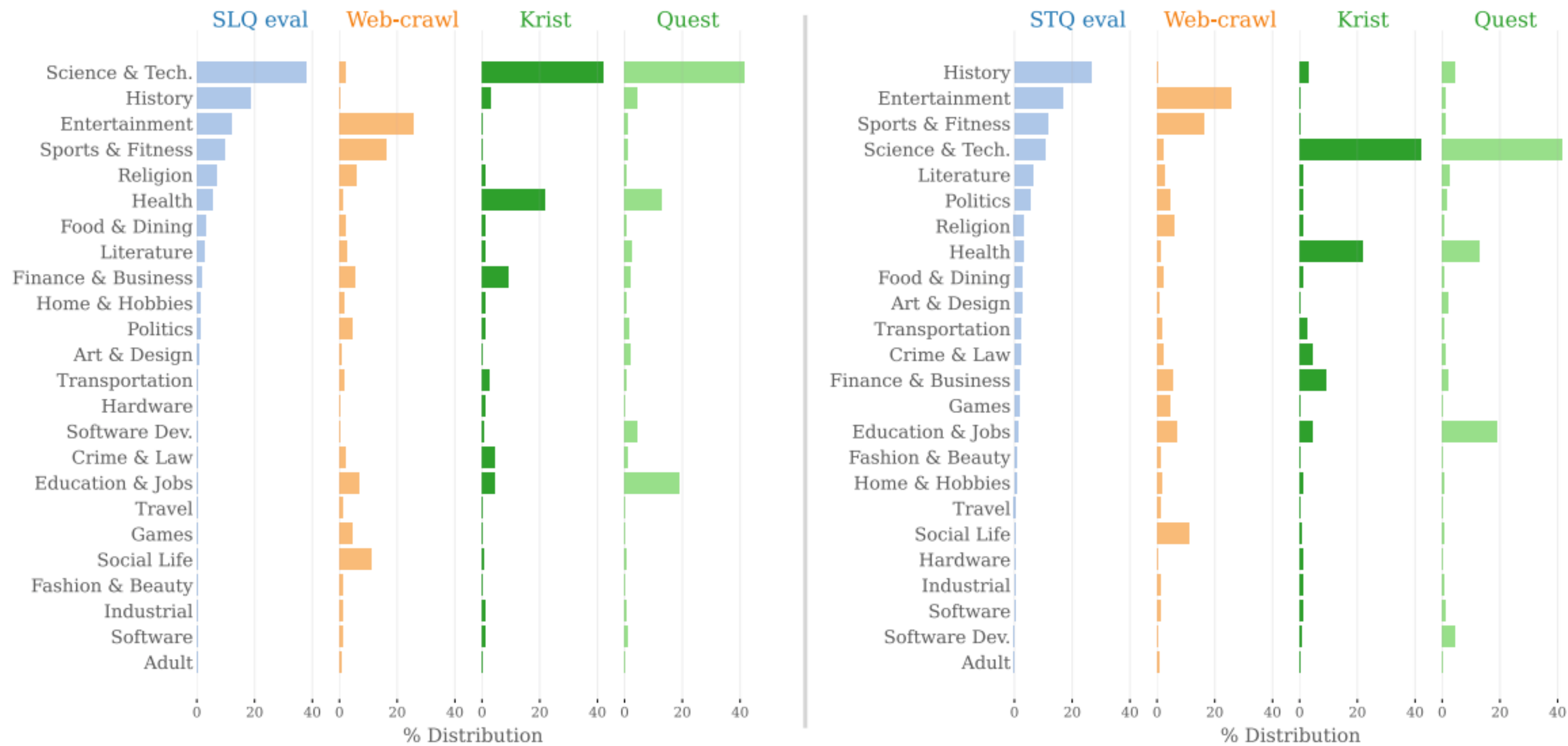
Result: Our data-centric findings also improve post-trained checkpoints

Pretrain ckpt	Text quality ¹		Audio Quality ²				
	spoken-alpaca	noisy-alpaca	Eval 1	Eval 2	Eval 3	Eval 4	Eval 5
coarse	42.6	45.2	37.4 (17.2)	33.3 (24.1)	34.3 (18.1)	37.0 (16.3)	38.8 (16.9)
fine	44.3	47.3	39.9 (18.5)	33.8 (23.7)	36.4 (11.6)	38.0 (16.9)	41.9 (20.7)
fine + syn	47.4	48.8	41.1 (17.1)	36.6 (23.1)	40.1 (18.7)	39.4 (16.9)	39.3 (16.8)

Why do our data interventions help?

Reason 1: Synthetic data boosts domain coverage!

Distributions of topic domains in our evaluation and training datasets



Why do our data interventions help?

Reason 2: Improved modality alignment!

