

ivrit.ai - Techincal AMA Session

Jan 2026

ivrit.ai - What? Why? Who?



Yair Lifshitz



Yanir Marmor

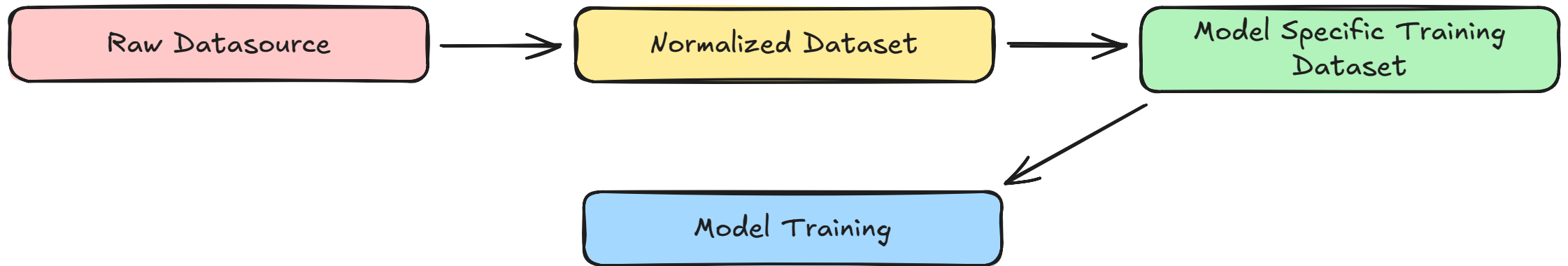


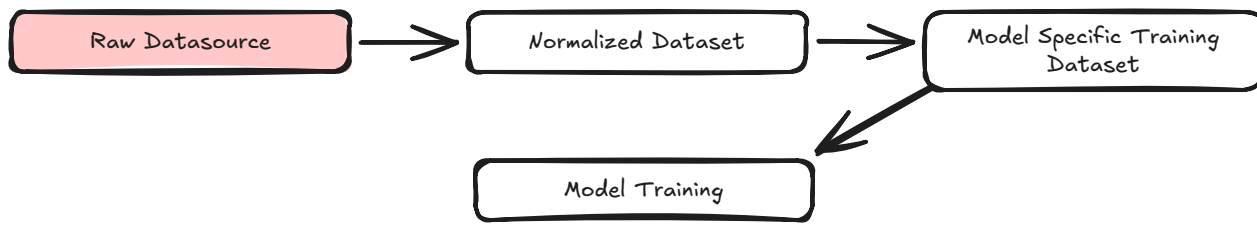
Kinneret Misgav



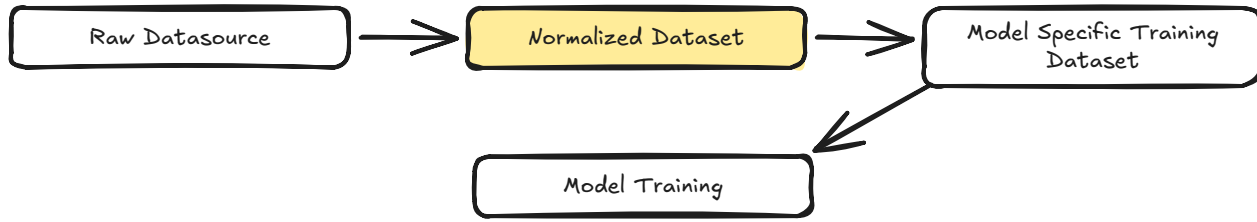
Yoad Snapir

Data -> Models -> Services





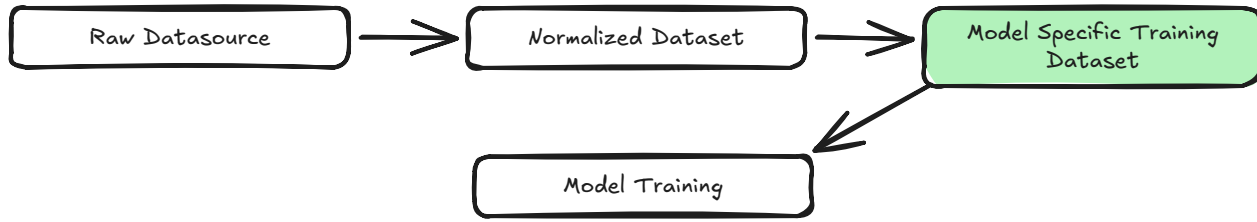
- Audio quality, format, and contents
- Text structure & format
- Existing text-audio timing
- Accuracy of text for the recorded audio
- Test split data selection strategy



- Decisions
- Audio representation as text
- Expected text symbols in ASR output
- Automate as much as possible as early as possible
- Get best alignment possible
- Also helps in quality assesment

Normalization code in the "ivrit.ai"
repo under the ivrit.ai GH
account.

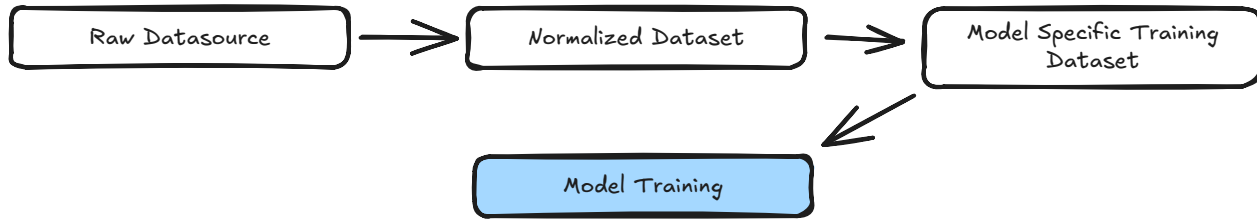




- Decisions
 - Length of Samples
 - Minimal quality to work with
 - Metadata to keep
-
- Produce Audio/Text training samples
 - Train/Dev Split

Dataset prep code in the "ASR Training" repo under the ivrit.ai GH account.





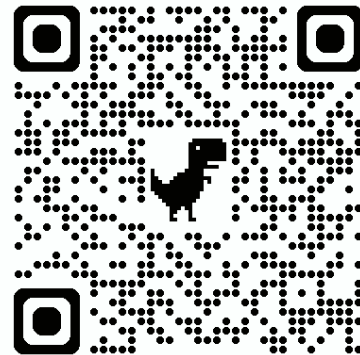
- Decisions
 - Base Model
 - Mixture of training data
 - Hyper Params
-
- Where can you train?
 - How long to train?
 - Evaluating training results
 - Choosing/Merging Checkpoints

Training Scripts GH Repo

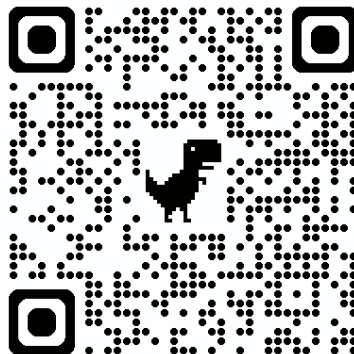




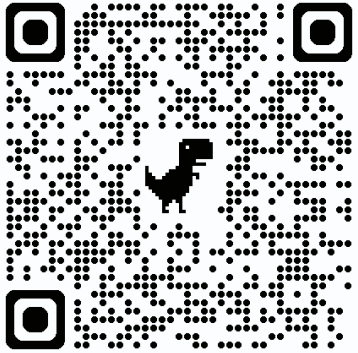
[ivrit.ai Huggingface](#)



[ivrit.ai Github](#)



[ivrit.ai Patreon](#)



ivrit.ai PyData Talk - Working on the Knesset dataset (2025)



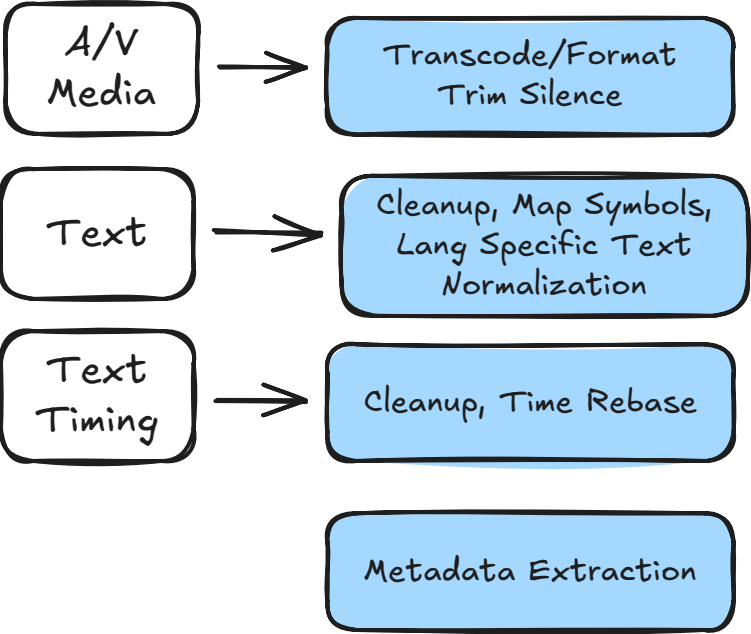
Blog Post - Training Whisper (2025)



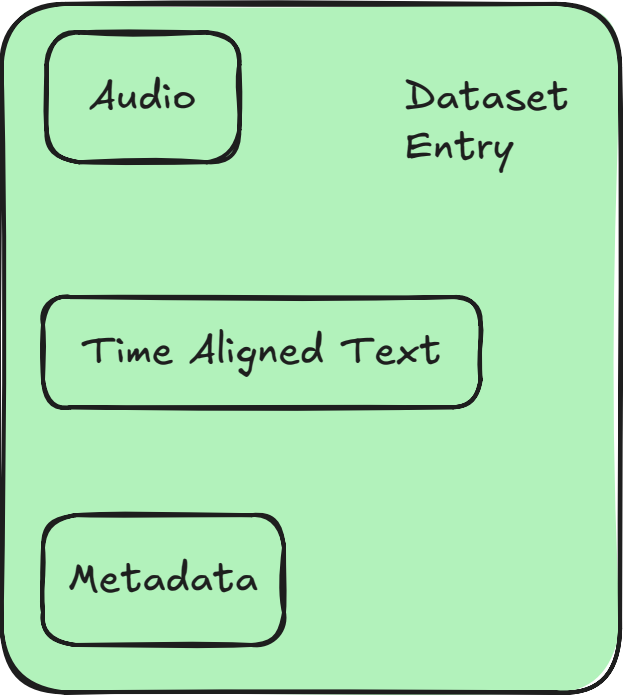
Blog Post - Long Form Transcription (2025)

Data Normalization

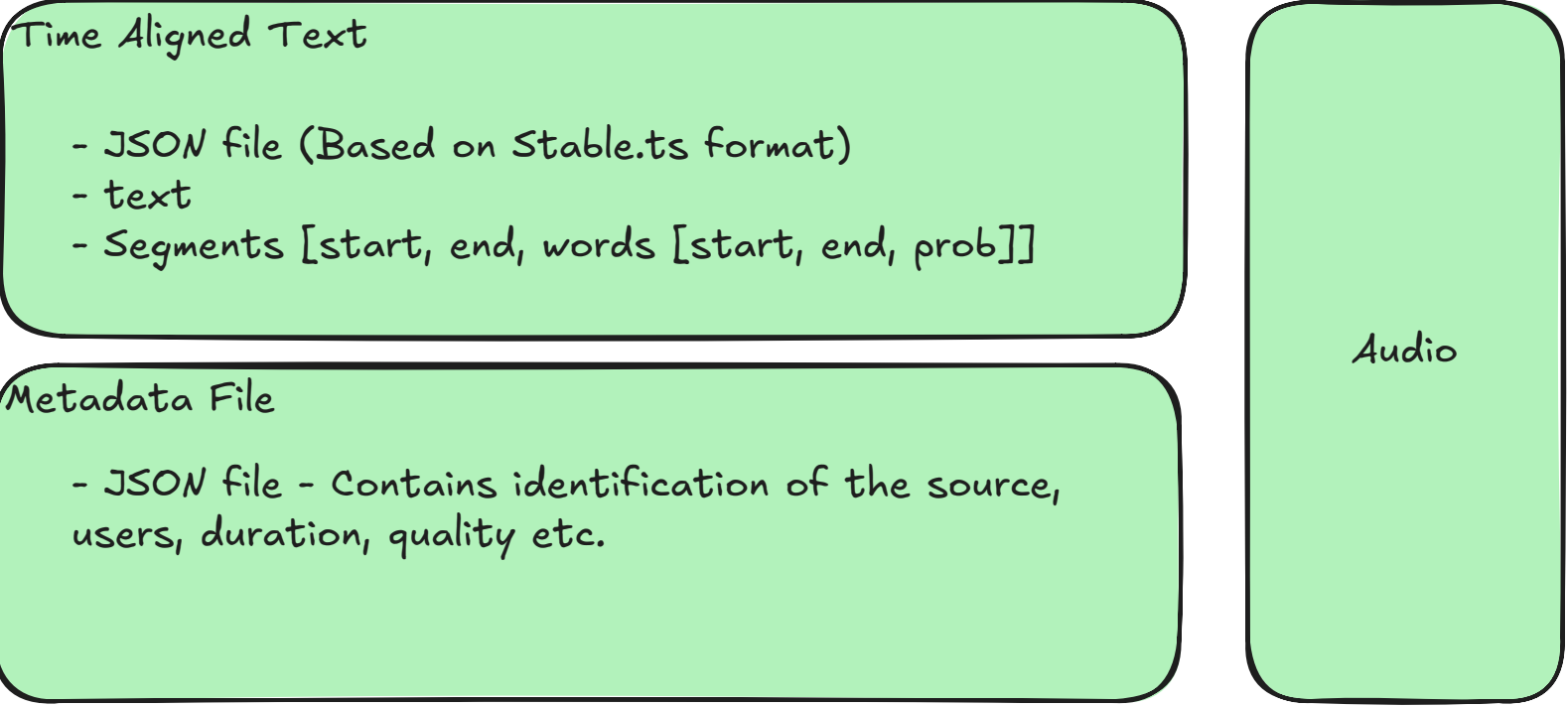
Raw Datasource



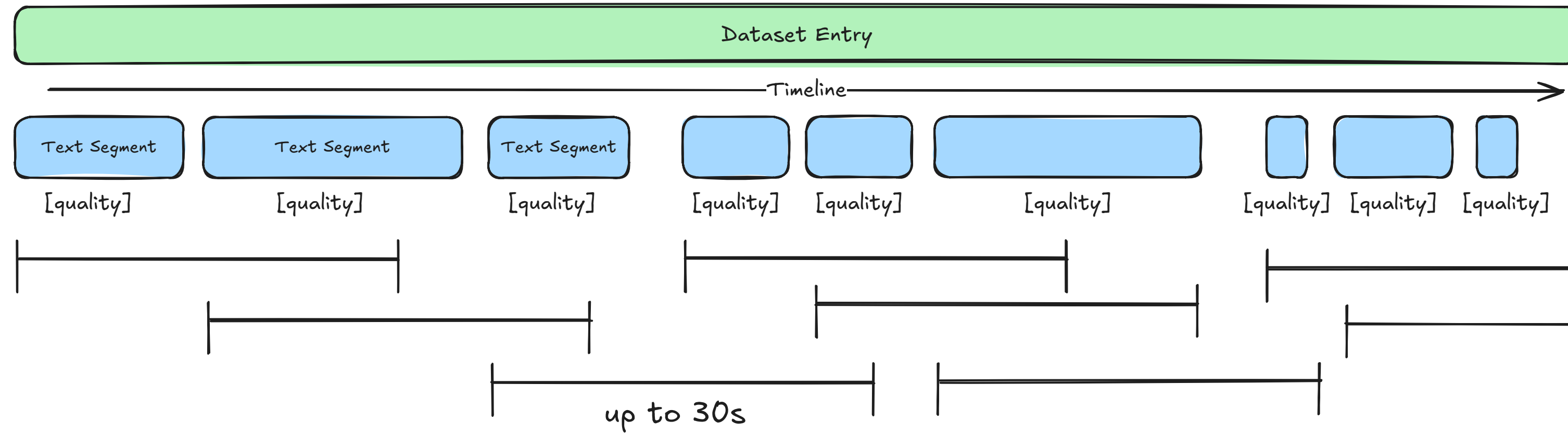
Normalized Dataset



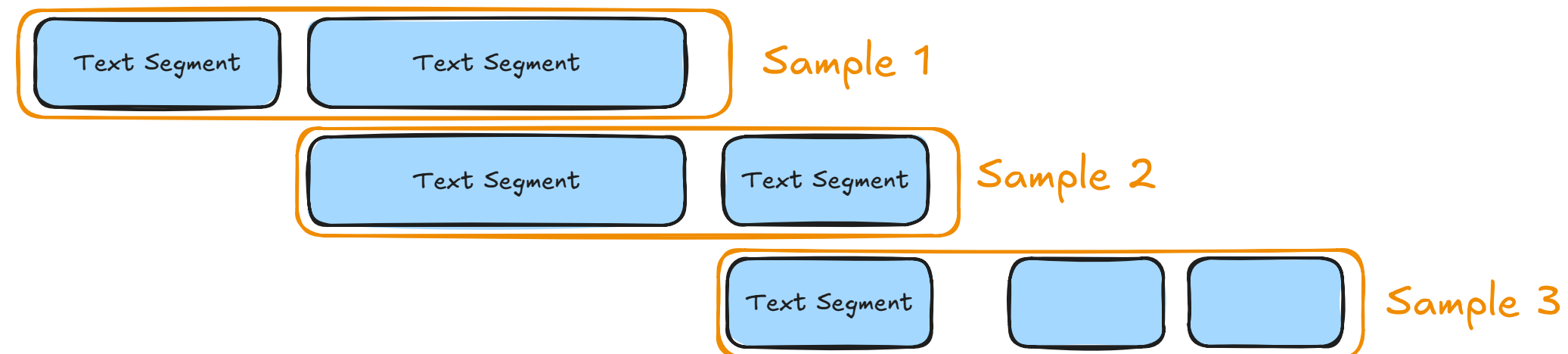
Normalized Dataset - Single Entry



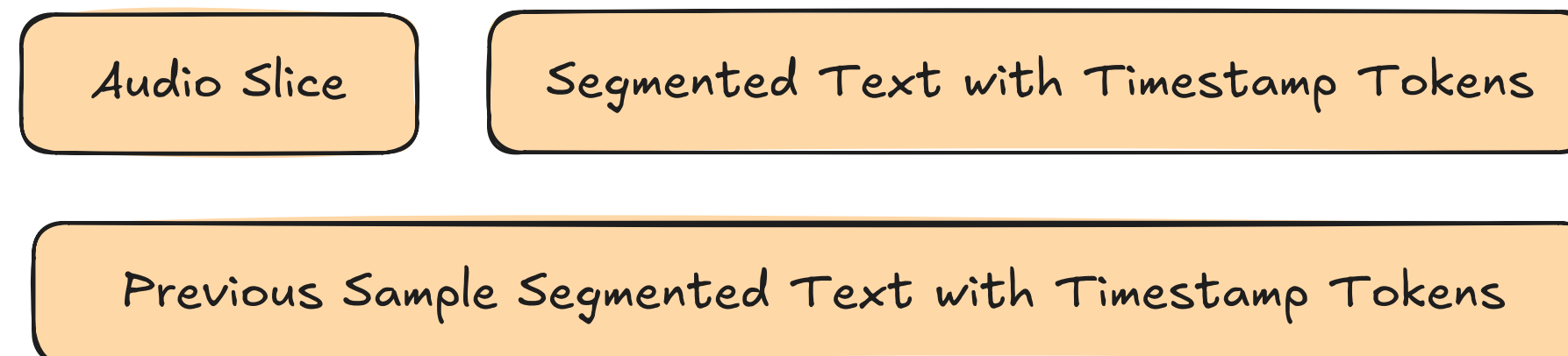
Training Dataset Creation



Splitting
To Samples



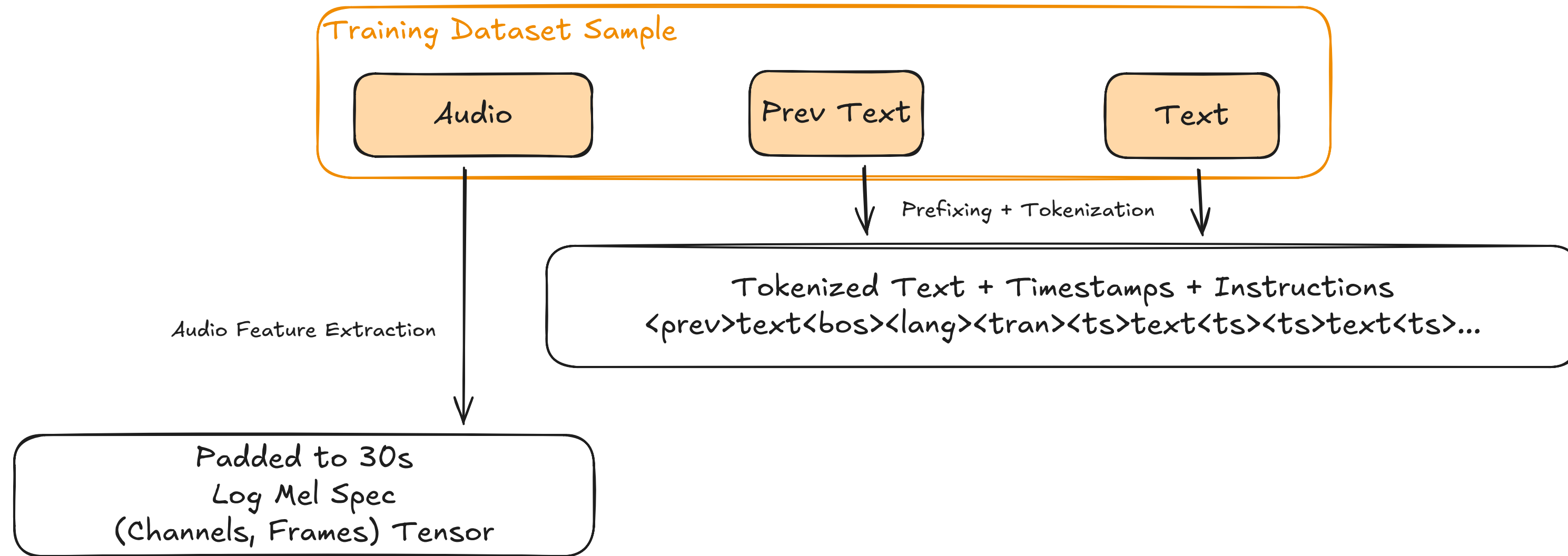
Training Dataset Sample



Notes

- Filtering by quality
- Here we also decide on the train/validation splits.

Training Dataset PreProcessing



Notes

- We remove padding before saving preprocessed datasets and re-add the padding back.
- Preprocessing is "model version" dependent V2 will not work with V3

ivrit.ai - Technical AMA Session

Questions...